DOCUMENT RESUME

ED 171 742 TM 008 770

TITLE	First California Conference on Educational Evaluation and Public Policy, 1976.
INSTITUTION	North Dakota Study Group on Evaluation, Grand Forks.
SPONS AGENCY	Rockefeller Bros. Fund, New York, N.Y.
PUB DATE	Feb 77
NOTE	76p.
AVAILABLE FROM	Center for Teaching and Learning, University of North Dakota, Grand Forks, North Dakota 58202 ($2.00)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Decision Making; *Educational Assessment; *Educational Policy; Elementary Secondary Education; *Evaluation; Evaluation Methods; *Evaluation Needs; Higher Education; *Opinions; Policy Formation; Program Evaluation; Social Values; Standardized Tests; Student Testing; *Testing Problems
IDENTIFIERS	*Evaluation Problems

ABSTRACT

A number of issues regarding educational evaluation and the problems associated with the practice of evaluation were discussed at a conference held in San Francisco, California, in 1976. Participants and persons submitting position papers included evaluators, administrators, teachers, educational researchers, parents, and students. Statements of participants and excerpts from position papers were presented on the following topics: the effect of evaluation on organizations; values of humanistic evaluators; student attitudes that standardized tests are unfair; student responses to testing; student evaluation needs; political influences on a school system's decision making; parents' confusion; alternatives to standardized testing; and teachers' perceptions of evaluator bias. A variety of problems associated with current program evaluations were discussed, and alternative methods were also proposed. A panel discussion was also held on cultural differences, white middle class values, the values of decision makers, social science research, tracking of students, and the concept of failure--as these topics relate to educational evaluation. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED171742

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM008 770

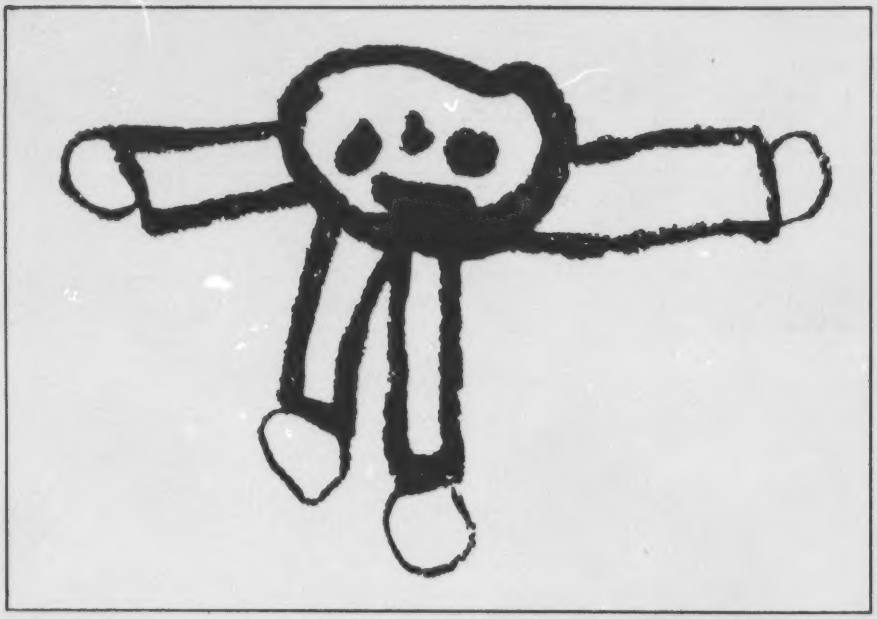
"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

D. Perrone

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

2

North Dakota Study Group on Evaluation



FIRST CALIFORNIA CONFERENCE ON EDUCATIONAL EVALUATION AND PUBLIC POLICY, 1976

University of North Dakota
Grand Forks, N.D. 58202
February 1977

Copyright © 1977 by Far West Laboratory

First published in 1976

North Dakota Study Group
on Evaluation, c/o Vito Perrone,
Center for Teaching & Learning
University of North Dakota
Grand Forks, N.D. 58201

Library of Congress Catalogue
Card Number: 77-701-72

Printed by University of
North Dakota Press

A grant from the Rockefeller Brothers Fund
makes possible publication of this series

Editor: Arthur Tobier

In November 1972, educators from several parts of the United States met at the University of North Dakota to discuss some common concerns about the narrow accountability ethos that had begun to dominate schools and to share what many believed to be more sensible means of both documenting and assessing children's learning. Subsequent meetings, much sharing of evaluation information, and financial and moral support from the Rockefeller Brothers Fund have all contributed to keeping together what is now called the North Dakota Study Group on Evaluation. A major goal of the Study Group, beyond support for individual participants and programs, is to provide materials for teachers, parents, school administrators and governmental decision-makers (within State Education Agencies and the U.S. Office of Education) that might encourage re-examination of a range of evaluation issues and perspectives about schools and schooling.

Towards this end, the Study Group has initiated a continuing series of monographs, of which this paper is one. Over time, the series will include material on, among other things, children's thinking, children's language, teacher support systems, inservice training, the school's relationship to the larger community. The intent is that these papers be taken not as final statements--a new ideology, but as working papers, written by people who are acting on, not just thinking about, these problems, whose implications need an active and considered response.

Vito Perrone, Dean
Center for Teaching & Learning,
University of North Dakota

Contents

	Introduction	1
1	Evaluation's Effect on Persons and Organizations	5
2	Evaluation Design and Theory	21
3	Evaluation and Public Policy	56

Introduction

Beginning in mid-1975, individuals and representatives of groups in the San Francisco Bay Area involved in educational evaluation began meeting with each other, having discovered that their misgivings about existing evaluation practices were commonly held. Many Bay Area evaluators felt, in particular, frustrated about the dehumanization of evaluation as expressed in the evaluation of national Title 1 and Follow Through programs. When they tried to voice these frustrations at various conferences, they were met with the absence of a professional ethic that could accomodate humanistic evaluation issues. On the other hand, they also discovered that the absence of a professional ethic didn't mean the absence of interest in the topic within the profession. At social hours and other informal gatherings, they found numerous colleagues who were troubled by the trend in evaluative practices toward the discounting of all educational outcomes save those easily quantified through standardized testing.

By the fall of that same year, Paul Chaffee of the California Council for the Humanities in Public Policy was actively meeting leading educational experts in the Bay Area to discern the policy issues surrounding evaluation that were then being debated. The Council, a state-based agency of the National Endowment for the Humanities, funds public programs of policy issues, asking grantees to include academic humanists in the dialogue. One of those with whom Chaffee met was Nick Rayder, an evaluation specialist with Far West Laboratory in San Francisco, who took issue with the evaluation of Follow Through.

Rayder and Bill Baker of the Alameda County School District, another critic of traditionalist evaluation practices, connected Chaffee to the Bay Area Consortium, composed of school district representatives from six counties and programs such as Berkeley Experimental School Project, Far West Laboratory and University of California (Berkeley) Field Services. The Consortium had been meeting monthly since the start of 1975 to discuss evaluation problems and theory. Its members were charging that educational evaluation, as it was then generally conceived, discounts the assessment of needs, personal consultations, reflective thought between reports, and contributions from humanistic scholarship.

The Consortium envisaged a conference extending and directly following the 1976 San Francisco-based conference of the American Educational Research Association (AERA), which

Acknowledgements, for their help on the monographs and in planning and conducting this conference, are due Joan Abrahamson, William Baker, Daniel Ben-Horin, Harene Noel Krenkel, Paul Chaffee, Elaine Boyce, Lee Conway, Marvin Gentz, John Newton, Garford Gordon, Michael Jang, Roger O'Connor, and Allan Abrahms.

the group found to be bypassing the real issues of educational evaluation.

A planning committee worked for six months putting the conference together; their framework stressed concern over the relationship of evaluation to funding and to public policy, a commitment to quality humanistic education, a conviction that innovative practices were being stifled by existing evaluation procedures, and a strong challenge to evaluation practices that failed to support pluralistic values in education. At the end of this planning period, the Council for the Humanities awarded the group a grant to help underwrite the conference. The Berkeley Experimental Schools Project agreed to act as a co-sponsor of the conference, providing substantial financial and in-kind support. A variety of groups, including the school districts of San Francisco, Richmond and Alameda County, University of California Field Services, and the San Francisco Classroom Teachers Association, provided the in-kind services. Finally, A San Francisco artist named Joan Abrahamson, acting as project coordinator, provided the continuity that held the planning and preparation together.

In order to move ahead within a conference framework, the planning body decided on the following organizational principles:

1. The conference should include concerned professionals from a range of advocacy positions in order to establish a rigorous context in which to consider some emotionally-charged issues and in order to convey to professional evaluators that the movement for evaluation reform is at least widespread (if not yet clearly focused).

2. Concerned individuals from the full variety of evaluation-impacted life and career roles should be included, and the atmosphere of the conference should be informal enough to permit each individual to feel qualified as an "expert" about his or her own evaluation-related experience. The conference should not be dominated by evaluation professionals.

3. The conference's format should reflect an appreciation of pluralistic values, should be flexible, and should be experience-based in order to convey a sense of urgency around issues.

4. The conference participants should feel that resources and expertise sufficient to make a major impact on evaluation procedures were present at the conference, and that the question to be faced was, "What can we do?" rather than "What can we make others do?"

With those principles in mind, the planners sent invitations to selected parents, students, teachers, evaluators, public policy-makers and humanists from backgrounds other than education; in short--to the evaluating, the eval-

uated, and those with a vantage point permitting a fresh and humanizing look at the entire situation. As points of departure for the conference, 29 key questions were drawn up, as follows:

1. What is the meaning of evaluation?

2. What value premises underlie educational evaluation?

3. How adequately are educational experiences being evaluated?

4. Are evaluators participants in a conspiracy of silence?

5. To what degree does evaluation contribute to social fragmentation?

6. What are the potential destructive and alienating effects of evaluation?

7. How can evaluation reflect humanism and social pluralism?

8. Can evaluation work for communication and personal development?

9. How does evaluation affect the individual student?

10. How does evaluation affect the school-community?

11. Who controls evaluation?

12. What are the consequences of not participating in evaluation?

13. What is the relationship between evaluation and funding decisions?

14. How does evaluation relate to program and curriculum decisions?

15. How does evaluation affect alternative education?

16. How are evaluation instruments evolved?

17. How is choice of instrument made?

18. Who evaluates evaluators?

19. What information are we getting with present techniques?

20. What are the alternatives?

21. How can evaluation be conducted in the arts?

22. Are resources being allocated to evaluation at the expense of program?

23. Is self-evaluation possible?

24. What do administrators and other groups need to learn about evaluation?

25. How does evaluation contribute to or detract from the education of doctors, lawyers, the clergy, artists, teachers, etc.?

26. Why is evaluation necessary? From the point of view of teachers? parents? students? taxpayers? policy-makers? administrators? curriculum planners? society?

27. What are the biases of "objective" social scientists?

28. How can the technician and the humanist work together to create a new form of educational evaluation?

29. What kind of policy changes should be made regarding evaluation and how can this be done?

This monograph presents the information gathered at the conference, as well as some of the perspectives formed there. In particular, it reflects the concerns that the participants expressed in three substantive areas: evaluation's effects on people and organizations, concerns and issues involving the design and theory of evaluation, and the way evaluation impinges on public policy.

1

Evaluation's Effect on Persons and Organizations

Where there is funding there is an evaluator

JOHN HEMPHILL · Most of the questions that evaluators are attempting to provide information about are beyond their competence to do so. They just don't have the techniques, the control of process to answer those questions, yet they never admit it. I have yet to see a proposal with money behind it that we couldn't find someone in the evaluation community who professes to be able to handle that problem. Yet I am fully convinced that many of those questions cannot be answered. *Statement during panel discussion*

John Hemphill is director of the Far West Laboratory for Educational Research and Development in San Francisco.

Organizational Consequences of Evaluation

GUY BENVENISTE · How does evaluation affect educational organizations? This paper provides a typology of consequences.

1. *Impact on Uncertainty.* In the immediate short run, evaluation sharply increases organizational uncertainty. Individuals in the organization are obviously concerned when outside evaluators who can affect funding come to look around. They are concerned when they realize they may be blamed for whatever is going on. Their natural tendency, therefore, is to retreat to more cautious strategies --to be sure the evaluators cannot blame them; they stick even more by the rules and avoid risks.

2. *Deterioration of Commitment to Innovation.* Innovators are...in a tenuous position within organizations for the simple reason that organizations do not reward innovators. Organizations only reward successful innovations into successes or failures, therefore into rewards or punishments. Each time evaluations are negative, they increase the anomie of the innovators. Each time they are positive, they simply say "what they should say." In other words, outside evaluators are perceived only as messengers of bad news and, as such, erode internal commitment to innovation.

3. *Deterioration of Professional Image and Norms.* Outside evaluation implies that there exists a meta-profession that can assess other professions. In social services such as education, where professional identity is weak, outside evaluation tends to further weaken professional image and norms...

Not all outside evaluations are going to weaken internal professional norms. We can conceive of professional evaluations that do, in fact, reinforce these norms. For example, if evaluations can be a shared learning experience, they will result in increased professional consciousness. But as long as evaluations are perceived as threatening, as long as they have other organizational consequences such as rewards and punishments, it is difficult to see how outside evaluators can avoid a deterioration of internal professional norms.

4. *Homogenization of Services.* One of the *worst* consequences of the use of benchmarks is that it results in conscious attempts by all service organizations to sharply reduce differences between units...

Those schools or districts who score high are fearful of scoring less high the next time around so they become very cautious; those who score low are fearful of the consequences of the low scores, and in a climate of fear, they are unable to improve. In short, the impact is negative across the board.

5. *Falsification of Measurements.* This is obvious if not well documented. The consequence is further anomie. The study of organizations in this country has paid little attention to the effects of corruption on organizations, but corruption can have very severe repercussions, particularly on professionals whose training does not prepare them for such behavior...

The problem with falsification is twofold. First, we know even less about what works and what fails; second, we have created a new kind of false profession, composed of individuals who really do not care at all about education but care only to look good in evaluations, and that is very serious, indeed.

6. *Reduction of Learning from Experience.* Last, but not least, evaluation should be part of the search for knowledge, part of learning. But in its organizational context, that is, in the context of the fear of punishment which accompanies outside evaluations--it is doubtful that much learning takes place.

One can easily describe the scenarios and stereotypes into which evaluation can degenerate. For the evaluators, the members of the organization being evaluated are unimportant, overcommitted, foolish, poorly informed, or clearly incapable. For the evaluatee, the evaluators are arrogant, wasteful of time and resources, committed only to studying instead of doing, uncommitted to 'real' education. This is not a good learning situation.

Moreover, evaluation takes time--time which the practitioners may feel is being diverted away from the task at hand. It is easy for conflict to arise between these two groups, and even when the evaluators think they have the enthusiastic support and understanding of the evaluatees, they should think twice and look again...

Excerpted from position paper

Guy Benveniste lectures in education, University of California (Berkeley).

How to be a naughty evaluator

JACK L. SLOAN · What are the fundamental characteristics of the institutions which have sunk so many of us in a vast sea of gloom and despair? One thing about schools and mental health programs that strikes me as fundamental is that they are hierarchical. As in the patriarchal family, allegiance is owed to the person (almost always a man) in charge, the big Daddy, the director, the principal or superintendent.

In the family, the Daddy enforces his power with the promise, if not acts, of physical violence. In our institutions the enforcing mechanisms are more subtle. Often they are fiscal. Often they include shared or delegated authority and high social status. The wellspring of the institutional Daddy's support is the funding source. In fact, the source of the money often *demands* that a big Daddy exist so somebody is accountable for the money and services, real or sham, that the money is supposed to buy ...

Many of you are highly credentialed and highly paid members of some Daddy's family. Often your Daddy will trust you to make some 'expert' judgments about what is important to find out about in and around your institution. Why not start finding out what the people in your community, those who use your services, want? Find out what makes them happy. Find out what they want that you might have on hand. Find out what they want that you might be able to get and when you find out, TELL IT! Publish locally, use the media, get the word out.

Also, STOP! Stop talking about your clients in terms that stigmatize them and turn them against each other and make some of them into freaks. It doesn't matter that you don't think a minimally neurologically impaired child is not a freak. It does matter that the other kids on his or her block do think the child is a freak. Stop labeling! Start reorganizing your evaluations to fit a system where each client is an individual who deserves an entirely custom-made individual treatment to fit that client's specifications.

What if your Daddy finds out what you are up to and doesn't like it? What if Daddy threatens to punish you for being too bold? What if he or she threatens to disown you? I can't say when you should give up or when you should get out. I have made enough trouble to end four or five careers and still I linger on. Sometimes being naughty is fun. I even quit once but my need to be a self-righteous loud mouth lured me back... *Excerpted from position paper*

Jack Sloan is research associate, Westside Community Mental Health Center, San Francisco.

The need for painful confrontation

HERBERT KOHL · Humanists usually care to do the best, most sensitive job they can, and to tolerate the values of all the people they work with. Humanistic evaluators usually come up with sensitive measures of feeling and commitment,

as well as with interesting judgments about process. However, humanistic evaluation has a major soft spot--there are areas of life in which there are conflicts that will not be resolved by people sitting together and communicating and sharing feelings and respecting each other and listening to each other. People with power do not like to give it up, and people who are poor and oppressed have to get power and command resources. Agreement on feeling is not the issue when one is dealing with conditions of oppression. Those in control and those who are their subjects must do more than agree that everyone has feelings, listen to each other, and get to know each other better. Those whose power is beyond their needs and whose power is based on the control of other people's lives have to give up that power or have it taken away. Those without power have to find strength and organization so that they will be able not merely to endure oppression but overcome it. In order to change things, one must be more than humanistic, more than caring and sharing of one's feelings. There are greater risks that evaluators and teachers and administrators and all people who are serious about eliminating oppression have to take than getting their feelings hurt. An evaluator has a double role--to work for a community that he or she chooses, and to fight for the values he or she has. If this puts one in opposition, it should be no surprise. The idea of an evaluator, one who works only to see what values are embodied in the works of others, but who has no values of his or her own and functions on no values, is absurd and usually a mask for the more depressing fact that the evaluator works solely for those who control the system, those who try to raise their values to the level of objective truth.

Excerpted from position paper

Herbert Kohl is author of *Reading: How To* and *On Teaching*, and an alternative schools teacher, Berkeley.

A student's view of standardized tests

COLLEEN WONG · I think evaluating students by standard tests is ineffective, unjust, and dehumanizing.

Third, sixth, eighth, tenth, and twelfth grade students have to take standardized tests. They are composed of vocabulary, reading comprehension, and mathematics.

These tests are undesirable, archaic, and faulty for the following reasons: Teachers often don't cover or emphasize the same materials. Students aren't evaluated, teachers are.

I question the value of the vocabulary portion. Checking off one slot is as easy for the student who guesses correctly as for the student who knows the answer.

> "What did you mark for the synonym of 'confirmed'?
> "I checked off 'denied,' but I was guessing."
> "I guessed too, but I marked letter 'c,' 'verified.' I hope it's right."

This conversation occurred after testing in my

eighth grade year.

Even if one scores in the 99th percentile on the vocabulary and reading comprehension parts, the student may not be adequate in English. The ability to creatively write an essay with smoothness and harmony is not measured. Neither is the ability to write good, brief stories that meet standards of journalism.

Supposedly, the math portion's objective is to determine the examinee's mathematical ability. However, some problems are described in English. Students who may be math geniuses may not be particularly adept in English. They can't understand the problems, much less get correct solutions. They might happen to guess correctly, though.

Also, what if a student doesn't feel well because of a family argument, a personal problem, oversleeping, having lost an important game the day before, or sickness? Too bad!

Some students can't sit still for several hours. Their attention span is just not that long. Some students get so bored that they just check anything after a while.

Selective colleges and universities attempt to utilize the standard tests to predict the future success or lack of it for its applicants and entrants. Are motivation, perseverance, experience, maturity, or other traits measured that are important to success? Can colleges afford such a misjudgment?

Standardized tests are not only ineffective; they are also unjust.

Students studying in different school districts and different schools have different skills. The recipients of an academic education generally have different capabilities from students of vocational schools.

The reading comprehension content of the standard test I took in the eighth and tenth grades is unfair. One story relates to traveling in Europe and another concerns the special features of a newly marketed car.

Don't such topics discriminate against the ghetto child?

The booklet asked where Copenhagen was. Why not ask the location of Djakarta or Madagascar?

Why not ask us the innovations of a newly marketed sewing machine?

What about the junior and senior high students who are stuck because they didn't learn the basics in elementary school?

Some students decide late in high school that they desire to attend an institution of higher learning. Standard tests may disqualify them from scholarships. Don't their willingness and ambition count for anything?

Ineffective and unjust. What more can standardized tests be?

Dehumanizing.

We have been misinformed that the test scores are a reflection of our worth as human beings. Actually the numbers only reveal the degree to which we have been prepared.

The tests don't help any. They are more damaging

than constructive. Students labeled 'high achievers' don't help those labeled 'dumbbells.' The 'flunkees' are separated into 'remedial' classes that are instructed by teachers who have low expectations of their 'backward' students.

In the next evaluation, they will do even more poorly because they haven't been helped and have become disheartened. They won't ever be prepared for the test. They aren't the culprits, however. No one motivated or assisted them.

Students planning to major in English who fail the math part will also be hurt. The math scores look bad on the records.

In my eighth grade year, the test results came out. Rumors spread regarding so-and-so's flunking and so-and-so's scoring high. The school counselors had asked some students to remove the adhesive score tags and place them onto the student files. They not only removed the tags; they read them. The whole affair was very upsetting.

Some students didn't and still don't know that files are being kept on them. They also don't know that test results are kept permanently. I ask: are our scores shown to teachers, to prospective employers, or to anyone else? Is the student's permission required? Is the student even told that his or her scores are being studied? If a computer makes an error, can the wrong be righted?

We are people. Please treat us as such. Dehumanization is out. Injustice is un-American. Inefficiency is hardly a progressive ideal.

Standardized tests are a waste of time, effort, and money. *Position statement*

Colleen Wong is a high school student in San Francisco.

Learning how to 'win' the evaluation game: a student's account

DAVID SCHWARZCHILD. Somewhere in my childhood I got the notion that I was not good enough, that I hadn't pleased my parents. It was a terse self-evaluation. For the greater part of my childhood, I under-valued myself; I felt that I stood in unequal stead with my contemporaries.

Evaluation is like a prehistoric bird casting its looming shadow everywhere. It is like an octopus flailing its tentacles into every nook and cranny. It is unavoidable. Whether in the form of the FBI, the CIA, or in peer groups, one has to be a hermit to avoid the tentacles.

I strove to be better than the other kids to compensate for my worries of inadequacy. Through whichever channels were available, I would charge forth, not giving an inch, disputing my inequities at the drop of a hat.

Evaluation is a part of the American way. Competition in our society is very keen, trends are towards bigger and better products. We are constantly being faced with advertisement, allowing a wide range of options to choose from, necessitating constant evaluation.

Education offered the golden opportunity in my bid

for proving myself worthy of acclaim. Elementary school was ideally rigged to handle my needs. I flourished in this system.

The purpose of evaluation in the education process is not clear. However, evaluation is never as prevalent as in the confines of the educational process.

In my elementary school I was able to exercise the need for approval and betterment without complication. I was highly motivated to do better than the other kids and the climate was perfect. The teachers conducted the classroom on the basis of competitiveness. They favored the achievers and placed them at the 'head of the class,' creating obvious room for comparison. I fared well by teacher's standards.

Evaluation in education is a blatant example of misuse of energy. The A-B-C-D-F grading system, which is prevalent throughout the learning institutions in America, channels away the student's vast reservoir of interest and motivation from learning, and engages him in the 'fight for an A' game. Worse, this rigid-type grading system gives the instructor the opportunity to set a very unhealthy, competitive learning environment because he has the student by the throat.

I gained recognition for my achievements, won approval from my teachers and, on the whole, relished my prosperity as well as my position. However, I could not help but sense that my relations with my friends were slipping.

Evaluation plays far too imposing a role in learning. The focal point tends to be the grade rather than the process of learning.

I felt uneasy about my success in the classroom, my favor with the teacher. I no longer enjoyed the subtle comparisons the teacher made. I felt uneasy about being elevated.

By inducing learning through competitiveness, there remains a residue of dissatisfaction, unhappiness. Everyone cannot gain the coveted 'A.' Consequently, learning leaves a bitter after-taste with many students--the ultimate infliction imposed by our learning institutions.

I was quite able to conveniently use my classroom, teacher, and the grading system to assert my 'superiority' over the other students. This, in itself, exemplifies the fact that many bad feelings can and will arise in a situation in which one person is elevated above another. The end effect is that learning in this type of situation loses all the beauty and meaning that it can have.

The American Heritage dictionary evaluates the meaning of the word *evaluation* as: "To ascertain or fix the value or worth of."

I question whether it is necessary to fix the worth of our students. *Position statement*

Why don't students evaluate?

Audience Participant: Why not think about an evaluation

process that is started and implemented by the youths themselves? You talk about a panel that is racially balanced--I don't see one youth on that panel. I don't see one idea coming out of the youths themselves. I really feel we should think about our representation in this room, because I don't see too many young people.

Maria Vargas: I would like to go back to what the students said about why don't the students evaluate. They don't because the way it is in the public schools, it's an issue of power. If students as soon as they enter the school system are given the ability to choose, they might not follow the teacher's directions, but if they don't follow the teacher's directions the teachers have no control. So what we create in the public schools is students who can follow orders and will accept our evaluations of them and so on up the line. And this is something that you can't change when you get to secondary school. I have tried that and the students don't want to bother--it's too much trouble to include themselves in decision-making. They are used to letting the teacher decide. So we have to start at the beginning and change that pattern. It's a very militaristic pattern.

Exchange during panel discussion

The evaluation students need

FRANK OPPENHEIMER · Students do need to know where they stand and what is expected of them. They need information that will tell them when to try harder, when to be satisfied and when to give up trying to learn some particular skill or field. However, the most effective way of satisfying this need does not come from grades or teacher evaluation; it comes from a combination of knowing what other students are doing and an awareness of the norms of adult achievement. A key to a pattern for satisfying this student need is suggested by the involvement of students in athletics or in music. In both instances, students can measure their progress with respect to their peers and the norms of adult performance without any intervention through teachers or institutional data gathering. In the case of athletics, there are sports magazines and television interviews which show not only the public athletic performance, but which also give detailed insight into training camps, employment opportunities and criteria, and some indication of the private life patterns of the athletes and their sponsors. In music, similar exposure is available, especially with popular music. The training and life routines of concert musicians is less available. Symphony rehearsals are rarely open to students and are rarely described in the public media except as fiction. Yet there is no reason that curious students could not be allowed to witness rehearsals and have individual contact with the musicians involved.

An instructor is certainly heavily involved in guiding the progress of students in both athletics or music.

He or she can point out particular deficiences and ways to overcome them and give advice concerning the reasonable expectations that students can have for themselves. But it is ultimately the proficiency of the student and his own interest that determines how far he or she will and can go. However, self and peer group evaluation can constitute the primary sources of evaluation at the outset. At a later stage these become supplemented with auditions and trials. In both instances, there is the practice of 'scouting' for talent and relatively little reliance on institutional certification.

The key pattern of athletics and music may conceivably eliminate the broader requirement of 'guaranteeing' the product, a requirement that currently so intricately and wastefully interweaves the process of schooling and education. Mechanisms must be devised and implemented which promote and allow two-way communication between learners and the active and productive aspect of society. The desirable breadth of education can then be a reflection of the breadth of a youth's contacts and experience, rather than an artificially superimposed 'welfare' curriculum. The establishment of mechanisms which encourage and permit such contacts would require rather fundamental changes in social values. In particular, we would have to believe that spending time with young people is not a waste of time; more generally we will have to learn to enjoy and take youth seriously. One of the multitudinous current demands on schools reflects precisely the opposite set of values. Schools are now asked to keep kids out of the way, to keep them off the streets. Most people become extremely nervous when they see a bunch of kids hanging around: "They should be in school."

It is true that in general one cannot have kids hanging around offices and stores and factories and film studios without paying some attention to them. But is this attention, this distraction from getting things done, a waste of time or is it the most valuable use of time we could conceive? In pre-industrial and rural societies, children were close to the activities of adults. Our society has invented and produced tools to do things cheaper, better, and faster. Why must these tools force us to spend less time with youth and to isolate youth by corralling them into schools? If young people are allowed to be around so that they could observe, work alongside, and eventually apprentice, then they could accumulate a personal experience record that would be far more reliable to a future employer or next step educational institution than a grade point average or a diploma.

Documentary television could also contribute much more effectively than it now does to showing the processes and the standards of human activity. How little is seen of what people do in factories, assembly plants, and warehouses! How does the manager of a hardware store decide what to stock? Does he look through catalogues, talk to his clerks, go to trade fairs? There are many fictional television shows of lawyers, doctors, policemen and air-

line stewardesses. But these are not very believable and they leave out so much of the drudgery and training and decision-making that is involved in all professions. Actually, novels do a better job than television in this regard, but television should be able to do much much more, with the skillful use of "theatre verite."

Excerpted from position paper

Frank Oppenheimer is a scientist and director of Exploratorium in San Francisco.

One parent's discontent with Berkeley public schools (impressions of an irrational educational community)

ALEX PAPILLON · As with most urban areas, Berkeley has a racial, cultural and economically heterogeneous community. Such heterogeneity places strong demands on the educational system. Different populations require both different kinds and intensities of educational programs. Where third world groups are political minorities, their needs can only be met (short of social disturbances) when the political majority makes a moral commitment to run schools commensurate with educational needs. Commitment must be evidenced in the areas of finances and support for the eradication of racism in the system.

In Berkeley, this kind of leadership is lacking. If anything, the leadership has exploited the atmosphere of racial hostility. Political capital is being made out of the elimination of programs designed to compensate for the effect of racism. Hiring policies resulting from the development of special programs, together with affirmative action programs, have been blamed for budgetary problems supposedly caused by overemployment. Experimental programs are being cut or eliminated under the argument that they failed or that they are too costly for the meager benefits derived. Such assertions are usually made with no (or inadequate) evaluation of the programs or their process of delivering education. Berkeley is totally devoid of an evaluation mechanism designed to provide a scholarly analysis of its programs and their educational process. This vacuum creates a ripe environment for racist exploitation of school resources. Further, in a situation already charged with racial hostilities, decisions made without objective and reliable data provide additional fuel for a climate of mistrust, thereby destroying any opportunity for collective participation in and support of the decision-making process...

Each group of parents, teachers, and school board is engaged in the political process of forming coalitions and attacking the opposing groups. Under such circumstances, the issue of academic skills is either fragmented or buried underneath the turmoil. A considerable amount of conflict is generated by each group blaming the other for educational failures; fear is created by the prospects of a child losing some educational resource; and uninformed views of what the system is doing and is capable of doing, act to keep the cauldron bubbling.

The parents, whose primary interest is that of aca-

demic skills, find themselves in the weakest position, although they are potentially the most powerful group of all the players. Because of the racial issue, the financial crisis and the teachers' movement, their concerns have been manipulated by the better organizational forces of the teachers, school board, and administrators. The consequence is that political fragmentation which leads to power moves will impact educational policy decisions by making them political decisions. In the ebb and flow of various crises, new coalitions are formed in a never-ending spiral of political decision-making. Thus the district stays in a perpetual state of chaos.

This result is fostered in part by the closed nature of the educational institution. Teachers, administrators, and the school board have shown a negative attitude towards parent power. Consequently, in order for parents to act collectively they must develop some consensus about educational priorities. Most parents agree that educational skills are the number one priority. After that, the consensus falls apart. In part, this is a result of forming opinions from prejudices and misinformation. A factual and objective view of what the system is presently doing, what the educational problem areas are, and what the system is capable of doing, would go a long way in establishing a basis for parents to collectively evaluate what their options are. Essentially, if the parents had data from a good evaluation, the move to parent power would take a giant step.

Of course, such data would not be a panacea for the divisions among parents. However, my experiences in organizing parents and participating in parent organizations suggest that reliable objective data goes a long way towards creating a rational dialogue between parents with differing views. Further, such dialogue is a meaningful first step towards compromise and consensus...

As a parent, the lack of trustworthy and relevant evaluation data frustrates my attempts to discover what the system will do for my children. In lieu of such information, I am forced to rely exclusively on a subjective evaluation. Given the turmoil in the system, one must conclude that chaos reigns supreme and that academic skills are secondary to the concerns of political power and district manipulation. Additionally, when evaluation is given a low priority my confidence in the system's ability to plan a comprehensive and meaningful program with grade continuity is severely shaken. I am compelled to rely on the political rhetoric to determine the priorities and emphasis of the system. When such rhetoric is dominated by what I consider to be racist concerns, worker-boss struggles, and unmanageable hostile fragmentations, as a responsible parent I must see the BUSD as an unhealthy environment for my children. I do not want to leave the impression that evaluation is the total solution to the problem nor that it is a perfect instrument. If people are intent to fight, manipulate, or remain in a state of hysteria, then no instrument is relevant. Fur-

thur, the instrument itself is imperfect due to its inability to measure such things as incompetancy and racial prejudice. What the instrument produces is in part reflected in what it is asked to do. This point is particularly relevant as evaluation relates to BUSD. It has been with consistent amazement that I have examined the confused, irrelevant, and distorted evaluations conducted by BUSD. Such things as continuity between grades, program pockets and effectiveness, and teacher affectivity are inadequately, if at all, examined. The relevance of variables, of race, economic status, and culture as they relate to program effectiveness are examined in a grossly superficial manner. It's incomprehensible how educational planning takes place within BUSD's heterogeneous community when program and demographic data is so totally lacking. If one was unknowledgeable about the low priority and lack of sophistication of the BUSD evaluation mechanism, healthy paranoia would suggest a conspiracy to defraud the public.

It is apparent that the necessary role of evaluation will not receive attention from BUSD. Given the racial climate, the financial crisis, and the teachers' movement, the lack of rational information by which to evaluate the academic competency of BUSD is simply too much to tolerate. *Excerpted from position paper*

Alex Papillon is chairperson of the Parent Advisory Committee (PAC) of Project Follow Through, Berkeley Public Schools.

How evaluation begets confusion for a parent

OBIE COOPER · There are three levels of evaluation currently in the school district. It seems to me that each level is busy pointing the finger at the other levels.

1. As a parent, one becomes very confused as to what is happening in the district.
2. Activities from one level aren't known about by the other level.
3. It becomes a full time job to sift through all the information to find out the *real* facts.
4. Parents who *think* they are involved are confused about the goals of all the different programs.
5. One set of evaluation procedures should be applied to all district programs.
6. The office of evaluation should be set up similar to business services. All information should be located in one place.
7. The dissemination of results should be a set procedure.

The evaluation should involve two tasks: (a) determine the layout, or overall plan, of a site or of each stage of the history, and (b) establish a chronology, or time sequence, of things to find. *Position statement*

Obie Cooper is a parent and ex-president of the Berkeley (Calif.) Teachers Association.

Teachers and parents attempt to find an alternative to standardized testing

JANE HORII · PAULA VELOSKI · LINDA WILGUS · For a number of years, we as educators have been concerned about the value

Overheard during the conference

"Teachers don't *have* to do testing in their classrooms."

"In California you do!"

"The good teachers *I* see break the rules."

"I am feeling frustrated. My real questions aren't being answered. How can parents go into a school, and evaluate what's happening in the classroom, and make changes in the environment for their children that work?"

"I am upset about it but what can I do. There is not too much I can say."

"We talked too much, we professional evaluators. I would have liked to have heard about other people's concerns."

and effect of standardized testing in our schools. Naturally, as primary level teachers, our attention has been focused on the specific needs of that age group. However, in the course of pursuing the testing issue, we have discovered far broader implications--not only for students, parents, and teachers at all levels, but for anyone concerned with the whole process of federal and state funding of public schools as well.

Our active involvement with the standardized testing process began in the spring of 1973 when, as K-1-2 teachers, we were faced with the annual chore of administering standardized tests of reading achievement (TOBY for kindergarten and the Cooperative Primary Test for grades 1 and 2) for the purpose of identifying those children eligible to participate in the ESEA Title I program. In the past years, each of us had been troubled by the anxiety and sense of failure that these tests produced in our students and by the doubtful validity of the test scores relative to actual classroom performance.

In hopeful desperation, we joined together and approached our on-site administrators for some alternative to the established testing process. Through their contacts with our school district's office of research, we learned that parents have the right to request that their children be exempted from a given test. This discovery led to the parental request process: that has been the basis of our 'non-testing' program. Since very few parents were aware even of the existence or schedule of standardized testing, much less of their rights regarding student exemption, we took the responsibility for informing them. (Throughout this process of parental involvement, we have emphasized that it is the parent's decision alone that can exempt a child from testing, and we have attempted to present the school district's reasons for testing as well as our reasons for opposing it).

That first year's effort was a very 'spur of the moment' operation, relying primarily on telephoned invitations to parents to come in to view the test and make a decision regarding it. Even on such short notice, the parental response was rewarding and a number of children were exempted from the testing. In 1974 and 1975, when we were able to plan ahead, discuss the issue at parent conferences, and inform parents in a more organized manner, the response was correspondingly greater; so much so, in fact, that most of the children in our classes were exempted from taking the test.

For the first three 'seasons' of our testing exemption program, the school district's compensatory education office allowed schools to substitute teacher judgment for test scores as the basis for identification and placement of students in the on site ESEA/Title I program. In the fall of 1975, however, we learned that district regulations regarding testing and funding had been changed so that standardized test scores were the only valid criterion for Title I/S.B. 90 eligibility (the only exception being kindergarten, where no test was to be administered; hence teacher judgment was to be accepted). The elimina-

tion of teacher judgment as an alternative means of placement would make children who were exempt from testing ineligible for the Title I/S.B 90 program. This ineligibility could, in turn, cause a reduction in the total number of Title I/S.B. 90 participants and hence a reduction in funds available to our school. It became imperative that we meet with compensatory education staff persons prior to the testing period to discuss the changes in procedure and to investigate any other possible alternatives to the standardized testing process...

The result of our writing directly to the state, combined with our principal's persistent appeals to the school district for clear guidelines, was to finally have written commitment from compensatory education and the district office of research to meet with us and a body of interested parents and staff on December 3, 1975 at Dudley Stone School.

The specific informational outcome of the meeting was provided primarily by Ms. Margie Baker of the compensatory education office. In brief, she confirmed that the district was no longer able to accept the teacher judgment criterion, that standardized test scores were the only valid criterion for funding, that while total school funds might be reduced if fewer students were eligible, they would not be cut off completely, and that it was difficult to say which school agency had the responsibility for informing parents. The less tangible but more substantial effect of the meeting was to mobilize a number of parents to pursue the issue further on their own. The result of their involvement was a school-wide parent/teacher meeting held at Dudley Stone on the evening of January 12, 1976. Representatives of the various agencies involved in the testing process were invited, as well as Dr. Dean McKennon of the NEA, who spoke of that organization's research on the validity of standardized testing, in general, and the CTBS, in particular.

The January meeting had the very important outcome of securing from School Superintendent Dr. Robert Alioto the commitment from the district to inform all district parents in advance of any standardized testing and to invite them to preview at their child's school any district-wide test prior to its administration. Additionally, we learned from Mr. Phil Daro of the State Department of Education that there is a waiver process, available through district application to the state, for *identifying* Title I/S.B. 90 participants by some means other than testing. Finally, we learned from Dr. Alioto that it might be possible to choose another, more acceptable test than the CTBS for use in our school...

Obviously, as teachers, our principal motivation in this issue has been our concern with children and their education. We resist evaluation means that we find dehumanizing, traumatizing, rewarding of failure and invalid in terms of evaluating the student relevant to this individual development and learning situation. Further, we question a system that makes the teacher accountable for

a child's learning but considers that teacher incapable of judging which children are in need of the special help available through state and federal programs.

Our alternative is a simple one--probably too simple and too humanistic to be acceptable to the very complicated agencies involved in evaluating public education. We would make the teacher, as the professional who works directly with the child, accountable for all aspects of that child's learning, from diagnosis to prescription to evaluation--all based on each child's individual situation and needs. The teacher's accountability would be measured by a review of his or her progress relative to his or her stated goals and objectives for each child.

Granted this system of the teacher evaluating each child individually would not produce the reams of tidy statistics provided by standardized testing, but it would have the distinct advantage of being relevant, personal, and valid. *Excerpted from position paper*

Jane Horii, *Paula Veloski*, and *Linda Wilgus* are teachers at Dudley Stone Elementary School, San Francisco.

Teachers question evaluation

MARIA VARGAS · JOAN RODRIGUEZ · What is an evaluator supposed to do? Assist me as a teacher to help students learn--not just reading and arithmetic but about themselves. Not the least is to help students learn how to evaluate situations which confront them so they can make choices which are suitable to them.

What are some of the problems that I have encountered with evaluation and with evaluators? First, there is an immediate class difference between the evaluators, myself the teacher, and the students. The evaluators are paid more and they have had to spend longer time at institutions of higher learning. Therefore they are considered more of an authority than I am. Yet, can they spend weeks in the traditional enclosure of a classroom with thirty students and make something happen between those students? How many of them were unsuccessful teachers who decided to move "ahead"? If the students are from poor or minority families, it is immediately clear that the evaluators are considered more important than they, the students, or their families. How many upper middle class prep schools get evaluated since they send their students to Harvard or Wellesley, the evaluators' measure of success?

Who are the evaluators? They are people who have been willing to become part of the educational establishment in order to get their credentials. They are mainly white and they certainly have become part of the middle class if they weren't before. They had to acquiesce to the status quo in order to become evaluators. They are part of the club that school administrators, school board members, school counselors, and even some school teachers belong to. They must support each other. The problems are easier to interpret if they are seen as the students' and teachers' problems, not the system's problems. How

can they deal with the irrelevance, the racism, and the sexism of the students' experience in the school? After all, they too sat through all those lectures in order to get where they are. After all, they probably are racist and sexist, too--subtly so, as blatant racism and sexism are passé.

These evaluators have been trained in objectivity and random sampling. Do they ever ask the students what their values and goals are? Are these given equal weight with the goals and values of the establishment and the evaluators?

Who hires the evaluators? Administrators and Board of Education directors. Is it significant in Berkeley that the secondary school children of the members of the Board of Education are in the alternative school which is Berkeley High's answer to a prep school? Is it important that if the evaluators don't tell the administrators and board members what they want to hear, which reaffirms their values and goals, they might not recommend them to other administrators and board members at national conferences?

Who are the students? They are from a multiracial society, although the evaluators might not be. They are from lower socio-economic classes, as a rule, in inner city schools. Let us take a sample group in California--Chicanos. As has been well documented by the Office of Civil Rights, the students from this culture do not find Chicanos or Latinos in positions of power in the school system or in evaluation. In fact, their teachers, administrators, and evaluators see their most important job as moving these students into the 'better culture' or else keeping them happily in their place low on the socio-economic scale. If students resist because they have already identified themselves with the mainstream's definition that they are inferior, they obligingly become failures. If they express their resentment, they become dangers to society. If they see the discrepancy between the realities of their home life and the expectations that teachers are giving them about the society of equal opportunity, they might call the teachers liars. Even evaluators know that this is insubordination.

Why is it that students don't learn to evaluate? They can evaluate their teachers, their schools, and their society. But then they might start making choices which would not submit them to the control of these teachers, schools, and society. *Position statement*

Maria Vargas is a teacher in the Experimental Schools Project in Berkeley.

Joan Rodriguez is a Berkeley parent.

2

Evaluation Design and Theory

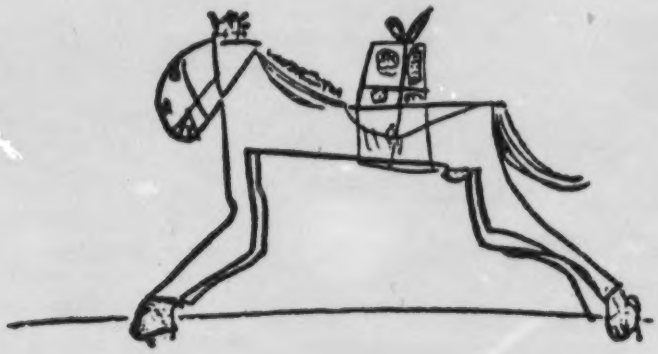
Methodological and ethical problems of research in early childhood education

"Something you see in your sleep is a...(a) dream (b) fairy (c) wish (d) dread."
Iowa Test, grades 1-3

NICK RAYDER · The absurdity of what is being done in the Follow Through evaluation was brought home to me by my daughter's drawings, and I thought it appropriate to show them here. My daughter is 5 years old and is crazy about horses. They enjoy a large part of her thought and creative talent. She drew some horses I would like to show you and discuss, as they relate to the 'horse race' paradigm.

Here's a horse that might do exceptionally well in a horse race. It's slim and peppy and could get a good jump on the pack. It probably runs short races best and may have trouble with a race over six furlongs.

Here is another horse. It has long legs and looks strong. From the feather on its legs you can tell it's probably a Clydesdale. It could pull heavy loads and might overcome obstacles that would be insurmountable for horse#1. Speed is not one of this horse's characteristics, but where brute strength and endurance were required this horse would probably do very well.



This next horse likes to graze in the fields, to romp around, and to deliver gifts to others. This horse likes to be touched and feels good with others around. This horse certainly doesn't look competitive and, if he had his choice, would not race against others. This horse seems independent but in touch with himself. He is a self starter and would probably never be found in a mechanical starting gate.

Above is another horse, with her foal. The mother/foal interaction is more important for both of these horses than horseracing.

The 'horses' in the Head Start Follow Through Planned Variation Program differ as widely as the horses in these pictures. They are models that have developed from different experience sets, extend from different value premises and reflect different educational theories.

As you are aware, the Follow Through sponsoring organizations were selected for their advanced work in education and in particular their work in translating educational theory into practice. The guidelines given them were to work with school districts (teachers, parents, children, administrators) to develop early childhood education programs to satisfy the needs and concerns of these groups. As expected, the Bank Street sponsor developed an early education model consistent with Bank Street philosophy and underpinning, and the Educational Development Center sponsor developed its model to reflect their own style of 'open education.' The group at Pittsburg University, Becker-Engleman at the University of Oregon and Don Bushell at the University of Kansas developed models that reflect a more structured, reinforcement-geared approach. Others applied the theories of Piaget, and so on. Still others focused on the home and on parent/child interaction. My own work has been with the Responsive Education Program, which has concentrated on training and re-educating school personnel to be more responsive to the needs of children, and more sensitive to the special needs of those from different family backgrounds.

The Planned Variation experiment has played fast and loose with research methodology. The original charge to the Follow Through sponsors was to develop programs that matched their own orientations with the needs and concerns of the individual school districts. The method of evaluation imposed on their efforts, however, does not take into account the tremendous variations in goals,

methods of implementation, and special local school conditions. In many respects, the national evaluation of Follow Through ignored the fact that Follow Through was designed to be innovative and to do something different than the usual school program. The evaluation paradigm imposed was that of a horse race, despite the fact that the 'horses' were neither racing nor even on the same track.

Usually educational researchers delight in variation since the goal of research is to explain it, and without variation cause and effect could not be explained. But in the Head Start/Follow Through planned variation, the large *variations* of programs being studied have defeated the evaluation paradigm. Let us just review some of the ways that the Head Start/Follow Through models differ among themselves.

Head Start/Follow Through models vary in their objectives:

- -- The focus for some is to change teacher behavior and institutions, for others it is to change the child.
- -- Some models concentrate on changing child achievement levels; others concentrate on improving the level and quality of parent involvement and building a functional partnership between parents and schools.
- -- (Etc. ...)

Head Start/Follow Through models differ in their testing procedures, and on how they view issues in measuring child development:

- -- Some models tend to accept existing standardized instruments as satisfactory for determining program effectiveness; others regard these instruments as measuring surface level knowledge of standard-type things and search for a wide variety of alternative ways to measure cognitive processes.
- -- Some models consider that instruments such as the Metropolitan Achievement Test (MAT) are appropriate to document program outcomes; others feel that the MAT is a poor way to assess knowledge because this type of instrument is unfair to children with different socio-economic and cultural backgrounds.
- -- (Etc. ...)

Head Start/Follow Through models differ on how test scores are used:

- -- Some use the test scores as an indicator of performance; others would observe the child while testing and note the behavior of the child as the indicator of performance.
- -- (Etc. ...)

Head Start/Follow Through models differ on how Head Start/Follow Through is being evaluated:

-- Some feel that, considering the developmental nature of Follow Through, all evaluation of Follow Through is formative evaluation; others feel summative data can be collected on Follow Through.
-- (Etc. ...)

Within this reality, and the reality that the initial legislation of Head Start/Follow Through planned variation was to develop--alternative, service-oriented educational models to treat the needs of urban poor--the use of the pre-test/post-test, experimental group/control group 'horse race' paradigm just won't work...

These methodological issues and problems inevitably raise ethical concerns. First of all, we should consider the enormity of the Follow Through project. Almost ten years of effort by hundreds of top-level educational developers and innovators in early childhood education have resulted in really fine educational programs that work. Teachers, parents, and administrators in hundreds of school districts across the country have cooperated in implementing these programs for hundreds of thousands of children. Hundreds of millions of federal dollars have gone into this effort--for development, implementation, and evaluation: over $25 million has been spent on evaluation alone. But more importantly, people's lives have been touched; children have received a broad range of medical, nutritional, social, and psychological services, as well as a more enriched educational experience.

The problem here is that we are looking at a social and educational project and it is being evaluated according to a design and methodology that does not reflect the comprehensiveness, the diversity, and the true nature of the program. Educational outcomes cannot be measured in accountacy terms alone.

Mick Ryder is director of evaluation for the Responsive Education Program at Far West Laboratory for Educational Research and Development in San Francisco.

Let me just bring up one small example of the absurdity that is involved here. One Follow Through community is a Native American reservation. The evaluation plan resulted in matching this community with an educational program based on token-reinforcement...

Excerpted from position paper

An evaluation of an evaluation

JOSEPH FILLERUP · I appreciate the opportunity to receive and respond to the draft copy of Volume 3, ABT Report, on Cohort II children. Our staff has spent several days together reviewing the materials and preparing comments...

If this document is to be one of the sole documents for decision-making concerning the Follow Through Project, then it is woefully inadequate in providing a comprehensive picture of the Follow Through evaluation effort. The audience for such a document would be primarily of two different types. One audience would be the decision

makers at the funding and program level. An audience of this type typically wants an evaluation which will present a complete description of context, methods and outcomes in the most concise and accurate presentation possible. This report does not meet the needs of this type of audience. The context surrounding the data is inadequately described and the methods are described in such a way as to give the impression that they are the only appropriate ways to analyze the data. Disadvantages of the methods used are mentioned only in a global frame with references to other sources. This audience is either not going to be motivated to refer to these sources or, if so motivated, will not be able to interpret these articles without the aid of a technical consultant. The audience of this type will tend to react to graphic representations of the outcomes more than verbal descriptions of the outcomes. In this regard, the graphic representations of the data in this report present an extremely biased view of the outcomes, at least in regard to this particular sponsor.

The second type of audience of this report is the technician who will be attempting to respond to the methodological and technical adequacy of this report. For the technician, data which would be necessary to make rational decisions are lacking. Again, the technician is referred to other articles which will describe the advantages and disadvantages of the particular methods utilized. However, not enough data are presented for the technician to determine which of these advantages and disadvantages may exist within this particular data set. Only enough data are presented to raise suspicions...

At first glance it appears commendable that three attempts were made at deriving an adequate control group for each sponsor. However, the impression is given that one can have more confidence in the results if non-conflicting results are derived from these three different comparisons. The major problem with this conclusion is that if the same problems exist in all three matches, it is obvious that the same conclusions will be derived from each of those comparisons. In the case of this particular sponsor, this is the case. In none of the three comparisons which are attempted is an adequate match made between FT and NFT sites. When this inadequate matching is made then all procedures and methods which are described for the analysis for such data are open to justifiable criticism...

The third area (we are concerned with) is the inadequate description of the context surrounding the data collection. An example of this would be that in one of our sites, data were collected subsequent to a teachers' strike within the community. Without a description of this context, many interpretations can be drawn from the data which are biased, inaccurate, and invalid...

We are concerned with the inadequate description of of the problems associated with the methods of analysis. The material significance criteria capitalizes heavily

upon chance difference. The report states that the problem of non-random assignment to FT and NFT groups makes statistical comparisons suspect. However, the utilization of material differences as defined within this report are subject to many misinterpretations of differences which would not be significant if chance effects were taken into account. Most of the graphic representations of the FT and NFT comparisons within this model would present a different picture if chance differences were eliminated from representation...

Excerpted from a letter to the U.S. Office of Education

Joseph Fillerup is director of TEEM Follow Through in Tucson, Arizona.

An evaluator talks back

MICHAEL SCRIVEN · Just for fun, and because the chance so rarely arises, I am going to role-play the tough evaluator reacting to common defenses or criticisms coming from program managers, and what she/he would say (at least under his or her breath) in response to these comments. Of course, it would be bad practice to come on this strong with most students or clients; but setting up a tough adversary is a good way to make one face the problems in the quoted positions.

1. *External evaluation no doubt has its uses, but it's terribly intrusive and just too expensive--and all that money comes out of what we are really here for, delivering services.*

Responses:

(a) You are not 'really here' for delivering just any kind of service, but *quality* service. Good intentions do not a good program make, and either you have a quality-control system going, to see if the program is good, or you don't. If you don't you should be closed down: Why should you be funded for what may be, as far as you or anyone else knows, shoddy or *damaging* intervention? (Think of the great efforts of well-intentioned and often heroic missionaries to deliver salvation and alter cultures that were the equal of their own.) If you *do*, in a quality-control system, then do you really suppose you can objectively evaluate your own services? If so, you are incompetent; if not, you've already *got* external evaluation and your responses have made clear exactly why it's essential. Not because *it* is going to be 'purely objective' but because it's going to to be *more objective in certain respects* than a self-evaluation, especially if you pick your evaluator carefully.

(b) Calling external evaluation 'terribly intrusive' either means you were once bitten by a lead evaluator, in which case you should try again and don't hire one until the evaluation plan reduces intrusion to a bare minimum; *or* it means you can't tolerate even a minimum of intru-

sion, which means you haven't faced up to what accountability in the management of funds *means; or* it means you think your operation can't be improved, which is touching but if true, you owe it to the rest of the world to show us how you became the first to achieve this, and showing an evaluator is the *least* intrusive way to show the rest of the world.

(c) Calling evaluation 'just too expensive' means either that you have been 'taken' in the past, or that you don't insist that evaluations pay off for you , which shows you don't understand *management*. There *are* certain types of externally-mandated evaluations, done for accountability reasons, which may have relatively little pay-off for you , but those ones are paid for by the funding agency. Whenever you have some control over whether or not to put money into evaluation, as the comment we are discussing implies, then you should expect and require that the evaluation give you a very good return on your investment. An evaluation can pay for itself in two ways. First, it can identify program components that are *wasting* resources so that they can be cut, representing a direct cost saving which often far *more than covers the cost* of the evaluation. Second, the evaluation can identify components that are extremely cost-effective, where you can get *more results* for a little extra support than elsewhere, thereby generating a larger quantity of quality service for the same dollars, which is again a major pay-off for you and the client population (assuming that their welfare is in fact one of your values). Good evaluators will often suggest sources of funds that hadn't occurred to you; or of personnel you need; or ways of making substantial savings on copying or communications costs. It depends on the range of their experience and expertise. But you should expect that *on the average,* external evaluation should pay off in the same way as other management consulting services or office equipment. That is, it should improve your performance and/or reduce your costs. If you have a truly excellent program, it *won't* pay you, because there won't be much room for improvement. But even a person who feels very healthy should have an occasional check-up: the cost of being wrong is much *higher* than the cost of the check-up.

2. *Look, it's easy to show that we are meeting our program goals, which is exactly what we are funded to do; what's to evaluate?* What's to evaluate includes, amongst other things, whether:

(a) You're also producing unfortunate *side-effects,* for example on community property values (typical

half-way house problem), or on clients (methadone), or on staff (overwork).

(b) The program goals are still appropriate, or are appropriate for your clientele in *this* location. (For example, sex or race segregated experimental schools, separate special education classes, vocational education training for no-longer-existant jobs.)

(c) The cost is excessive; could the same be done for less, or more done for the same amount? Maybe the specific goals (objectives) were set too low.

3. *What we're doing is obviously better than anyone else in the same field, around here; and a client/staff survey shows very positive feelings towards the program. In short, we have a very high quality program, and that's not just our opinion, it's the opinion of some very distinguished consultants and observers who know this area.*

(a) "Better than anyone else" may still not be worth funding (electroshock therapy, totally unstructured classrooms are possible examples).

(b) Your consultants and local observers may easily have become co-opted, either because they're paid by you, admired by you, or related to you (do you refer clients to them, for example?).

(c) Even if the preceding queries were answered, there's a basic point which by now the reader should recognize as an automatic refutation of number 3, just as it was of number 2. It can be summed up in this slogan: *Quality Is Not Enough.* Programs always feel deeply offended by this assertion--and indeed, it's hard enough to deliver quality. But it's *not* enough. There is always competition for the program money and it's anti-social not to recognize that; the implication is that not only quality but also (i) the *severity* of the needs being served, (ii) the extent to which other programs are meeting these needs--that is, the *indispensability* of the program, (iii) the cost of the program must be considered; in effect, evaluated.

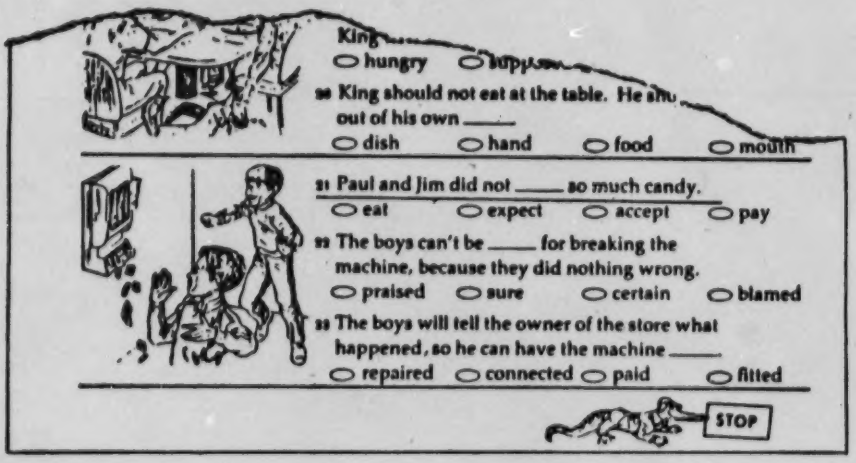
4. *No other program is doing what we're doing; there's nothing providing the unique combination of services we offer these clients.* Like the preceding comment (number 3), this one is often uttered in a hurt tone, as if uniqueness was a virtue in itself. In human service areas, from education to health, *uniqueness* or novelty is not a value *at all*, and it most certainly does not establish indispensability. At first, it looks as if "there's nothing providing the unique combination of services we offer these clients" means that the program is indispensable, is meeting a need no one else is meeting. But what (nearly always) turns out to be the case is that no one is offering this *combination* of services, though the important *components* of the combination *are* available else-

where. *Now* the decision is much more difficult. We have to ask whether the convenience of the single package is worth any extra components (multi-purpose agencies rarely perform as well on each dimension as the specialized ones). Uniqueness Is Not Enough Either...

Excerpted from position paper

Michael Scriven is professor of philosophy and education at the University of California (Berkeley).

CAN YOU ANSWER THIS?


King
○ hungry ○ happy
30 King should not eat at the table. He sho
out of his own ____
○ dish ○ hand ○ food ○ mouth
31 Paul and Jim did not ____ so much candy.
○ eat ○ expect ○ accept ○ pay
32 The boys can't be ____ for breaking the machine, because they did nothing wrong.
○ praised ○ sure ○ certain ○ blamed
33 The boys will tell the owner of the store what happened, so he can have the machine ____
○ repaired ○ connected ○ paid ○ fitted
STOP

Iowa Test of Basic Skills for Grade 1-3
(discussed in Conference Workgroup)

Passive bias and conflicts of responsibility in the evaluation-producing community

CLARK C. ABT · What might be done to reduce or eliminate (the) sources of bias and passive censorship? On the supply or production side, the findings, approach, and supporting analysis and data could be summarized in sufficiently brief form to maximize the communicative efficiency of the research. On the demand or consumer side, contracts could specify in advance support of the publication of all evaluation research results in professional journals of high repute to assure both high quality standards of reporting and evenhanded dissemination of findings, whether or not they support particular policy views.

The motivation of the producers of evaluation research might be directed to a more objective and uninhibited expression of results, provided such a practice was instituted through publication and professional journals and any losses of client confidentiality were avoided by previous client clearance. The motives of the evaluation producers should be to produce competent and honest re-

search, and open publication and the review and criticism of professional peers is more likely to encourage this than private publication in government reports of limited and selected distribution.

There are political, economic and scientific constraints operating on the unbiased and uncensored dissemination of evaluation research. Policy-making sponsors have a right to the first review of the findings of evaluation research, and although I am aware that this is controversial, I believe that they have the right to the privacy of these findings for at least a period of time sufficient for responding to the findings without embroilment in public controversy. In practice, this might mean a delay of a few months to at most two or three years in open dissemination--a delay entirely consistent with the usual operationally dictated delay in professional publications. These would seem to be the most important policy constraints.

Economic constraints on unbiased and uncensored dissemination are usually quite modest. For a $200,000 evaluation study less than 1 percent of the total cost of the research is needed to produce a good quality professional paper for journal publication. For the larger, $1,000,000 evaluations, perhaps 1 to 2 percent or $10,000 to $20,000 is needed to prepare a professional quality book-length manuscript. The economic support for this form of dissemination could easily be supported as part of the evaluation research funding, and a simple administrative requirement for 1 percent of the research funds to be thus expended is unlikely to have significant impact on the quality of the work done by the other 99 percent of the funds available. On the contrary, the quality and quantity per cost is likely to improve, if researchers know that they are writing for open publication and the esteem of their peers. Further, the editorial quality controls naturally exercised by professional journal and book publishers is also likely to raise the quality of the communication through the internalization of editorial benefits...

Those questions of what constitutes responsibility in developing and expressing the findings of evaluation research are further complicated by the issue of *what constitutes responsibility in communicating those findings.*

Should findings be communicated to all those who need to know? To all those who have a right to know? Or to those who have paid for, contracted for, and sponsored the evaluation research?...

The problem of deterrence to open communication of evaluation research findings by the threat of retaliation by the funding agency is not academic. In a recent private conversation with the cabinet officer and secretary of a major federal department, I asked him what would be the result of an evaluation researcher openly publishing the results of a study inimical to the policy aims of the sponsoring agency and without the agency's permission. He stated unequivocally that all contract and grant awards to the offending evaluation research would be reviewed at the highest level of the executive agency with the intent

of finding some good reason for not making awards of further research support. I asked him how often, in his experience as secretary of the department, he had actually had to resort to such measures to discipline a researcher, an he admitted that he could not remember actually ever having done so in several years. What this little anecdote suggests is that government officials may be quite ready to use their administrative power to enforce client loyalty, although in practice they rarely have to do so because the threat of doing so is sufficient to enforce the selective censorship involved.

In this situation, there is no real conflict of conscience for some evaluation researchers, who have no clear basis for knowing whether their loyalties in providing information should be given to the immediate agency sponsor of the research, the sponsor's sponsor (Executive Department), the Congress as the ultimate funding body, the public in general which Congress represents, or the scientific research community. Given this uncertainty, most evaluation researchers who survive tend to select the client and immediate sponsor of the evaluation research as the one owed the most immediate loyalty. Unfortunately, this situation reinforces the ability of the sponsoring agency to selectively bias and censor evaluation research results, lends support to the Congressional and public criticism of such bias and censorship, and may ultimately result in reducing continued support for evaluation research...

It would seem desirable to persuade agencies sponsoring evaluation research that their long-term interest would best be served by obtaining a high quality of unbiased and uncensored research, and that this could best be achieved by a greater degree of independence of the reporting of the evaluation research results from sponsor control.

This deterrent to unbiased and uncensored dissemination of evaluation research results could be removed by an explicit legislated requirement for open publication of all results within some reasonable period of time after research completion, retaining the right of the sponsoring agency to have the results first and privately and well in advance of open publication.

In practical terms, protecting the independence of evaluation research results and their dissemination from the perceived or actual control of the sponsors, could be achieved by the same requirements for professional publications of *all* evaluation research results previously suggested as a means for overcoming the problem of passive bias and censorship through selective dissemination.

Excerpted from position paper

Clark Abt is research manager, Abt Associates, Cambridge, Mass.

To test or not to test: evaluating how youngsters read and use language

JAMES MOFFETT · One conventional way for those outside the classroom to evaluate what happens inside is to in-

stitute periodic testing with instruments sanitized supposedly against any bias of the teacher or the school. For this purpose many commercial tests are put out both independently and as part of a curriculum package. They are not of course sanitized, because teachers whose job security is linked to their students' test scores may teach so closely to the test as to bias results heavily toward favorable scores. The only advantage of such evaluation is that student performances may be compared with those of other times and places--if that is an advantage.

Norm-Referenced: For most well-known standardized tests, this comparison occurs in two ways. First, they are usually 'norm-referenced,' which means that the score of anyone taking the test now is compared with the scores of some original 'normal' student population. So a student obtaining today a reading score indicating 'third grade, fourth month' is simply being scaled by the norms of that first population. Second, a student today may be compared with his contemporaries throughout a school district or, in most cases, the nation. The great weakness of norm-referencing, of course, lies buried back in the original 'normal' population. How was *its* normality determined? Is it normal for today also or only for then? But most of all, do norms established by performance of any population in public schools as we have known them do anything but set low standards? Norm-referencing reposes on what some student *did* do, under all the usual handicaps of conventional language teaching, not on what students *could* do under improved learning conditions. Teachers, parents, and administrators happy with 'third grade, fourth month' may be accepting a meaningless standard, possibly a very low standard that holds back many youngsters.

Furthermore, comparison itself should be challenged. It serves, of course, the immediate practical purpose of selecting out students for this group or that class or certifying some for admission and employment. In other words, the more schools operate by *limiting* membership or admission, by segregating and screening for their own or others' purposes, the more comparative evaluation seems to make sense. Parents need comparison only to the extent they are using their child to keep up with the Joneses. The student does not need comparison to 'know where he stands' because good learning processes always show him by feedback how well he is performing, so that his only reason for comparing his performance with others' would be to know where he stands in the eyes of adults manipulating his destiny. For your own diagnosing and counseling purpose, comparison among students has no value.

To ensure meeting minimums, schools are choosing more and more another type of test called 'criterion-referenced,' which measures students absolutely, against a fixed standard rather than relatively, against each other...

Consider the main purpose of criterion-referencing. It is not to distribute students against each other on a

curve. It is to find out which students or how many students can do certain tasks. The tasks tested for tend to set a floor, because they are selected as indicators that students and teachers are achieving *at least* such and such. Any test tasks have to gear themselves somehow to realistic expectations of what students may achieve. If this gearing is not built in by some kind of averaging of what students do in fact achieve--by morning--then it has to be accommodated another way, because schools can't administer tests on which too many students fail--for political reasons if for no other.

Criterion-referenced tests ensure that too many do not fail by including mostly very safe items. They focus on minimal standards. They are a pass-fail kind of test and assume that the large majority of students will pass. But how can they assume this without a prior score group? Obviously there is a kind of score group in the minds of the test makers, only it is not a particular population actually run through a particular test but rather a general notion of what most students have done, and can do, based on common school experience. Most children learn to master the long-vowel spellings--at least long enough to pass such a test, even if they never really read.

The only value of criterion-referenced tests is to cover one teacher for one year. It came to the fore in an age of mechanical accountability. It aims to get masses over a minimal threshold... In short, criterion-referencing differs not so much from norm-referencing as might appear at first blush, because both set low standards based on moving large masses a short way. This low center of gravity owes to the misguided practice of moving all students at once in the same way, of standardizing.

For true individualization the only relevant measure is the student against himself. If schools take each individual as far as he can go, charting experience year to year, they will accomplish manyfold what they do by standardizing...

The Alternative To Testing: Reading aloud is an alternative to standardized testing of reading. (For one systematic way of evaluating children's sight reading, see Kenneth Goodman's NCTE pamphlet (National Council of Teachers of English, Urbana, Illinois), ***Miscue Analysis.*** But most evaluation can be based on kinds of daily observation not involving you so directly.

Follow the principle that evaluation is an organic part of your everyday role, not a separate function done on special occasions. *Evaluation should not dictate, distort, or displace what is meant to measure.* The secret is to evaluate by means of valid learning activities without making students do additional activities only for testing purposes. The most efficient curriculum allows students to spend all their time learning. The more scorekeeping, the lower the score!

Evaluating by daily observation makes most sense, however, as a natural part of a classroom management that

arranges for individuals and working parties to be doing different things at the same time, as in a student-centered or open classroom, where the teacher can circulate to consult, counsel, and coach. The same passivity, paucity, and poverty of traditional classroom dynamics that make learning to read and write seem harder than they really are make evaluation seem like an inevitable parasite. The brute fact is that *ordinarily students don't do enough to provide the evaluator something to see.* But if students are constantly producing and receiving discourse in a great volume and variety, and if the teacher is freed from emceeing to circulate and observe, then good evaluation becomes possible without resorting to special activities that detract from learning and make students hate reading and writing...

Excerpted from position paper

James Moffett is editor of *Interaction*, a textbook series on language arts.

The science of watching and wondering

GEORGE F. HEIN · There is currently a debate in the field of educational evaluation between those who advocate reliance on standardized tests and those who insist that a broader approach, involving perceptions which are less susceptible to measurement, is necessary. A look at other fields shows that the same type of debate occurs widely.

Essentially the issue is between those who favor a 'reductionist' approach--one which seems to reduce problems to their smallest, and most scientifically analyzable, components--and those who favor a 'holistic' approach--one which is based on the belief that problems must be discussed in their natural context, as a whole.

The entire history of science is one of controversy between adherents of these differing approaches. The field of biology, for example, has been dominated for several decades by molecular biologists--reductionists to the core. However, within the last few years, there has been an upsurge of attention to the more holistic view represented by the likes of Goodall (observer of chimpanzee behavior) and Tinbergen (observer of herring gulls) who characterized his approach, in accepting the 1973 Nobel Prize for physiology and medicine, as 'watching and wondering.'

The common focus of reductionists' attacks on holists is that the latter are unscientific. In regard to this type of accusation, it is important to distinguish between holistic advocates of alternative scientific approaches and believers in knowledge gained exclusively through insight who argue against science totally.

In educational evaluation, as in other fields, our task is to blend the best of reductionist and holistic approaches. To this end, we need:

1. To continue to validate broader, more holistic approaches to evaluation, while not failing to concede the shortcomings of these approaches as they are perceived.

2. To defend alternative scientific approaches, while at the same time indicating how they differ from *non*scientific approaches.

3. To recognize that the superior sophistication of methodology currently possessed by the reductionist (test-oriented) approach to evaluation does *not* mean that this approach is therefore essentially more valid. There is a half century of experimentation behind evaluation testing, while the holistic approach to evaluation is a relatively new field. We must recognize the *potential* contribution of the 'watching and wondering' approach, given a similar investment of energy and research. *Abstract of position paper*

George Hein is former director of Follow Through at EDC, Newton, Mass. and currently dean, Lesley College, Graduate School of Education, Cambridge, Mass.

Deciding what is of value

HERBERT KOHL · Evaluation simply means deciding what is of value in a specific situation. However, given this simple definition, there are a number of questions that immediately come to mind.

1. *Who decides what is of value?* This is a major question for educational evaluators. There are parents, students, teachers, administrators, and school board members to be considered. In many communities the values of these groups are not the same. For example, in Berkeley there are major differences in the kind of education different groups of parents value for their children...Some parents value the freedom provided by libertarian education; others are skills-obsessed. Some students disagree with their parents about what is of value in the schools, and some school board members believe they know what is of value for all children...Add to this teachers who do *not* believe they should be evaluated but rather that the burden of evaluation should fall on the material used, and the burden of failure should be put on the student or the parent or the administration. Then throw in teachers who believe they serve the parents and have to be evaluated on the basis of what the parents want, and administrators who do not want to be evaluated so much as to control the evaluation process, and one gets a notion of the complexity and political nature of discerning who has the power to set values, and deciding who should have the power to set values.

I believe that neither teachers nor administrators nor social scientists should have the power to set values for other peoples' children. All of these people are workers who serve a community. If they disagree with the values of the communities they work in, they can try political education. They can try to teach and persuade and convince people to change their values. But they should not stomp on people and presume the right to set values regardless of what people believe and want for their children. One of the main reasons for so many schools becoming armed camps to protect the teachers and administrators

from students and parents (an insane situation) is because so-called professionals have presumed the right to set values for people, and therefore never found it necessary to ask or listen.

In a number of communities, there have been moves toward community control of schools and toward decentralization. These moves, though often co-opted by professionals, have usually originated in the desire of parents to take control of the definition of values *away from* presuming professionals. Parents have not for the most part wanted to become teachers or administrators so much as to decide and articulate what they want for their children, hire people who claim they can do what the parents want, and then hold these people responsible for what they do. An evaluator in this situation would be a consultant to the parents, a person with some knowledge and skill in assessing performance and helping parents set up criteria and methods for determining if what they hoped to have done was *in fact* done.

A number of communities have struggled to attain such community control. One group, the United Bronx Parents, provided assistance to a number of communities in New York where parents were taking control over schools and looking critically for the first time at what was happening to their children. With the parents, United Bronx Parents designed report cards to evaluate the schools--the schools were to be graded, *not* the children. The parents were the ones to decide what was of value...

2. *Who evaluates and what gets evaluated?* Parents can do evaluation themselves; they can farm the evaluation out to people they trust or they can delegate it to principals, to the staff, to students. A mechanism has to exist whereby parents and students can learn with a minimum of mystification about the few techniques and mechanisms of evaluation. For all the diversity of instruments and statistical techniques, the actual sophistication is not very great. Also one can understand exactly what an instrument does without necessarily being capable of running the results through a computer. There are times when one would think that professional evaluators have a stake in making their techniques look more complex than they actually are...

The judgment of taxpaying people, the basic sense that their children are not being treated well, is something evaluators have to respect and develop themselves. I do not believe that it is possible to design a sensitive evaluation instrument or an overall evaluation of a school or a district which will not produce results that are very similar to the judgments of parents and students. I also believe that it is possible to use even the most traditional instruments and come up with a confirmation of those judgments...

3. *Of what use is the evaluation to anybody?* An evaluation document should be a teaching document. It should reveal what is of value and be specific enough

about what is going wrong that it provides some hints about how to change things or at least about the source of the problem. The kind of evaluation of reading skills provided by a standardized test is a perfect example of an evaluation without any value to teacher or parent. It doesn't indicate anything specific that might be wrong, nor does it point toward what might be done. To be in the sixth grade and read on a 3.5 level is some sort of stigma. But 3.5 cannot be translated into anything beyond the fact that someone performed on a given day in a given way on a given test. *Excerpted from position paper*

School evaluators and the double-bind

BEN KLEINDORFER · Schools currently outfit youngsters for a hierarchically organized society where work is extrinsically motivated (graded). And the schools reflect that society in that they themselves are large, impersonal, extrinsically motivated institutions. Such organizations require abstraction and quantification in order to defend the way in which they cut up budgets. This is the purpose of quantitative evaluation.

Individualized qualitative evaluation can only take place when people know, live, and work next to each other; i.e., in small organizations.

Abstract evaluation gives rise to a feeling of meaninglessness, valuelessness, and anger in the evaluator and in the evaluee. Such evaluation involves separation, manipulation, arbitrarity, and is, necessarily, vulgar when dealing with learning.

Evaluators, therefore, are intrinsically placed in ambivalent positions. Perhaps evaluators can surmount this double-bind by recognizing it for what it is, and by doing whatever they can to democratize schools and their own working places. Out of this situation, more than anyone else, evaluators should derive motivation towards bringing about in our society a state of high, rich, and comprehensive learning rather than a merely narrow, mechanistic, uniform learning designed only to fit people into organizational niches.

It will be through an improved understanding and love of learning and education among equal nonsubservient citizens that evaluators will gain relief in their present positions.

Evaluators should spread the word that there is no scientific theory of learning, and that becoming an educated person will always be a mysterious, unmanageable, but encourageable process. *Position statement*

Ben Kleindorfer teaches management science at Pennsylvania State University.

CTBS 5TH GRADE

Down there between puffs of cloud in a brown-green pattern
is a micro-world where nothing appears to move.
There are people of course — they laid out the checkerboard squares —
and we saw them at first on the roads like a scatter of ants
which must have been cars. But now they are soundless,
invisible. This is all we really know,
a world the size of a plane, hung in the sky.

We know with the top of our minds (and the captain told us)
how many miles per hour this jet is going;
yet we seem held, unhurried, a part of the air
which is oddly solid beneath us. That waiting world
will grow into size and importance an hour from now,
where we shall feel proper concern for schools, and classes,
and teams, and matching shoe-colors, and washing the car,
and forget how the sky is wide
and the earth is small.

41 The world looks like a micro- (very small) world because

A nothing is moving
B the plane is so high
C the plane is so fast
D the sky is so wide

42 The "checkerboard squares" in line 3 are probably

F cars
G small farms
H rivers
J other planets

43 While in "a world the size of a plane," the poet is mostly aware of the

A problems at school and home
B schools, classes, and teams
C shoe-colors and cars
D wide sky and the small earth

44 The jet seems to be

F in a hurry to get back to earth
G growing in size
H part of the air, unhurried
J going 1,000 miles per hour

45 The best title for this poem is

A The Hanging Plane
B Jets
C The Quiet Plane
D Jet Flight

Evaluating the 'form of quality'

SISTER IRENE WOODWARD · We seem to have reached a place in our development as a society when we want to talk and hear about values and quality and ethical activity. This type of discussion is a move away or beyond our search for objectivity and for efforts to be nonjudgmental and value-free in our considerations of society.

We come to this need, however, without adequate tools for evaluation of the qualitative, esthetic, and ethical aspects of our lives. Surely, we have made tremendous strides in evaluation procedures during the recent past to measure, compare, and predict with regard to the quantifiable and, even, to apply those methods through statistics to sociological aspects of existence. This development in the tools of evaluation was a remarkable contribution to human thought and is not to be decried. Any misapplication of these procedures was just that, a misapplication and, at times, abuse, not to be confused with the sophistication of the procedures themselves. As we were developing such methods, we were either stunting our growth in the realm of the nonquantifiable or we were attempting to submit all human experiences to the same criteria. Hence, our position now.

The problem of judging and comparing those aspects of our lives which do not fit the mathematical mode has always been present. The ancient Greeks cherished and sought to evaluate and define 'beauty,' 'honor,' and the 'good life.' Medieval scholastics defined, divided, exemplified, and placed in hierarchical order 'justice,' 'art,' 'knowledge,' 'wisdom,' and on and on. Francis Bacon was certain that he could discover the 'form of quality.' Throughout history, most concerned thinkers realized that things must be evaluated according to some standards, that judgment is not simply each person's own interpretation; and, at the same time, that evaluation in the nonquantifiable realm is tenuous and complex.

Now, we realize ever more desperately that to develop new modes of evaluation is not simply an intellectual luxury but is absolutely essential to the future of our democratic society. We must be a group of people able to make value judgments in our businesses, in politics, in education, and we have not learned how. A course in legal ethics designed to help pass the bar is not what we need. A mode of education, of public discussion, and of widespread common concern is essential to a society based on the enlightened judgments and choices of individuals.

In order to move toward a basis for evaluation and a refinement of our ability to judge in a nonquantified mode, we must develop some common experiences, some models of excellence, and a vocabulary proper to our endeavor. Common experiences can arise out of conversations, articles, discussions, conferences, artistic productions, etc., in which the humanists, artists, and philosophers who have been in the wings until now can be given the opportunities to work together, investigate together, and

probe. Too often, in the past, such colloquia have been stifled because funds for such events were awarded only to grantees who could propose and evaluate according to criteria contrary to the very nature of such dialogue. Models of excellence can be developed through grants, awards, and encouragements to the genius, the gifted, the careful writer, thinker, ethician; and in a milieu which would foster excellence, both the models and the perceptive judges would be subtly developed. From these developments and discussion, new forms of speech and distinction of judgment would be refined, out of which could grow more sophisticated forms of evaluation in this other realm.

All of this may seem very far removed from evaluation and laboratories as we have known them. It is, and it isn't. The human ability to design a model against which to judge other forms is a constant. The kinds of models and the relations among qualities are different. But if we have been able to progress as far as we have in one mode, surely we have the resources to help ourselves strive for a kind of excellence and beauty and honor that we all deserve as human persons. *Position statement*

Sister Irene Woodward is president of Holy Names College in San Francisco.

The perceptive eye: toward the reformation of educational evaluation

ELLIOT W. EISNER · The major focus of this paper will be upon the description of a new form of evaluation that could have important consequences for the conduct of educational practice. In deciding to address myself to educational evaluation I in no way intend to back off from the perhaps more complicated problems involved in the design of curriculum. I focus upon evaluation because I believe that there is an intimate relationship between the assumptions and procedures we employ to assess educational effectiveness and the kinds of programs that we offer. Indeed, I would like to persuade you that the assumptions and procedures used in conventional forms of educational evaluation have, in the main, been parochial. They represent an extremely narrow conception of the way in which educational evaluation can be pursued...

What I would like to do is to suggest, perhaps not so much an alternative, but surely a needed supplement to the use of scientific procedures for describing, interpreting, and evaluating educational settings. I call this new, nonscientific approach to educational evaluation an approach that requires educational connoisseurship and educational criticism.

What I propose starts not with a scientific paradigm but with an artistic one. I start with the assumption that the improvement of education will result not so much by attempting to discover scientific methods that can be applied universally to classrooms throughout the land, or to individuals possessing particular personality characteristics, or to students coming from specific eth-

nic or class backgrounds, but rather by enabling teachers and others engaged in education to improve their ability to see and think about what they do...

It is an old truism that scientific studies in education are as often defined by the form of research one has learned to use as by the substantive problems one believes to be signiticant. Becoming familiar with correlation procedures too often leads simply to questions about what one can correlate; the existence of statistically reliable achievement tests too often leads to a conception of achievement that is educationally eviscerated. Our tools, as useful as they might be initially, often become our masters. Indeed, what it means to do any type of research at all in education is defined, stamped, sealed, and approved by utilizing particular premises and procedures. A brief excursion into the pages of the *American Educational Research Journal* will provide living testimony to the range of such premises and procedures. For example, during the past three volume years the AERJ has published over 100 articles. Of these, only three were nonstatistical in character.

Yet the range, richness and complexity of educational phenomena occurring within classrooms are wider than what can be measured. Some phenomena can only be rendered. It is this richness and this complexity to which educational connoisseurship addresses itself.

Imagine a ninth grade class in algebra. The school, in an upper middle class suburb, is one of those single story, try-hard modern school buildings that looks like the district ran out of funds before it could be finished. The classroom is a box-like environment, anonymous except for the sundry array of hand-carved initials that decorate the wooden desks that line it. Not even a nod is given to visual interest. The only visual art in it is a dog-eared poster of Smokey the Bear, reminding students to put out their campfires. The rest is blackboard, light green plaster walls, and small windows well above the level through which students can look out onto the world. Suddenly the buzzer pierces the atmosphere with a baritone voice that startles even the geraniums. To the students it is only a familiar reminder that class is ready to begin.

How do these tightly-zipped, vivacious youths enter the room? What expression do they wear on their faces? What does their posture, their pace, their eyes say about Algebra I to which they return each day at 10:05? How does the class begin? Is the procession of discourse one that stimulates, satisfies, is it one of dutiful routine, one of feigned enthusiasm that so many students of the well-to-do do so well?

How does the teacher respond to students? What does the tone of his voice say to those who do not understand? Do you detect the tinge of impatience that humiliates?

How is the class paced? Does it have a sense of movement and closure? Is competitiveness engendered by the subtle but pointed reminders of extra credit, of the

quiz the day after tomorrow, and of grades for college? What is the quality of the ideas and the analysis with which the students venture forth into the life of the mind? What is this teacher, in this setting, saying about algebra *and* about education to the students who meet him every day at 10:05?

These are some of the qualities, ideas, and practices to which an educational connoisseur might attend. But educational connoisseurship is not reserved for those outside of the teaching profession. Educational connoisseurship is to some degree practiced daily by educational practitioners. The teacher's ability, for example, to judge when children have had enough of art, math, reading, or 'free time' is a judgment made not by applying a theory of motivation or attention, but by recognizing the wide range of qualities that the children themselves display to those who have learned to see. Walk down any school corridor and peek through the window; an educational connoisseur can quickly discern important things about life in that classroom. Of course judgments, especially those made through windows from hallways, can be faulty. Yet the point remains. If one knows how to see what one looks at, a great deal of information--what Stephen Potter refers to as 'danda' in contrast to 'data'--can be secured. The teacher who cannot distinguish between the noise of children working and just plain noise has not yet developed a basic level of educational connoisseurship...

Connoisseurship when developed to a high degree provides a level of consciousness that makes intellectual clarity possible. Many teachers are confronted daily with prescriptions and demands from individuals outside the teaching profession that are intended to improve the quality of education within the schools. Many of these demands the teachers feel in their gut to be misguided or wrong-headed; the demands somehow fly in the face of what they feel to be possible in a classroom or in the best interests of children. Two examples should suffice. The pressures toward accountability defined in terms of specific operational objectives and precise measurement of outcomes are pressures that many teachers dislike. Their distaste for these pressures is not due to professional laziness, recalcitrance, or stupidity, but is due to the uneasy feeling that as rational as a means-ends concept of accountability appears to be, it doesn't quite fit the educational facts with which they live and work. Many teachers, if you ask them, are unable to state why they feel uneasy. They have a difficult time articulating what the flaws are in the often glib prescriptions that issue from state capitols and from major universities. Yet, the uneasiness is often, not always, but often, justified. Some objectives one cannot articulate, some goals one does not achieve by the end of the academic year, some insights are not measurable, some ends are not known until after the fact, some models of educational practice violate some visions of the learner and the

classroom. Many teachers have developed sufficient connoisseurship to feel that something is awry but have insufficient connoisseurship to provide a more adequate conceptualization of just what it is...

The end of criticism, unlike connoisseurship, is that of disclosure. Criticism applied to classroom phenomena is the art of saying just what it is that is going on in that setting. Take, for example, that mode of human performance called teaching. What is it that teachers do when they teach? How do they use themselves? How do they move? What level of tension, of affect, of spontaneity do they display? To what extent do they reveal themselves as person to the students with whom they work? Are they approachable? In what ways? What kinds of values, ideas, and covert messages do they emphasize? How, given questions such as these, can the qualities to which such questions guide us be disclosed? How can they be disclosed in a way that does not rob them of their vitality as experienced? Here the educational critic has a task similar to his counterpart dealing with live theater. The critic's task in each case is to provide a vivid rendering so that others might learn to see what transpired in that beehive of activity called the classroom. What the educational critic employs is a form of linguistic artistry replete with metaphor, contrast, redundancy, and emphasis that captures some aspect of the quality and character of educational life.

In this task, the educational critic does far more than describe behavior. A strictly behavioral description of what teachers do would not only avoid dealing with the intentions of the teacher, it would also describe in quantitative terms the number of behavioral moves made by the teacher. One such description goes like this:

> Launching is primarily the function of the teacher rather than the pupil. The teacher speaks 85.2 percent of all structuring components; in contrast, the pupil speaks only 10.2 percent. The range for the teacher is from a high of 96.8 percent in Class 6 to a low of 73.5 percent in Class 13, with a median of 86.2 percent. The range for the pupil, on the other hand, is from 23.6 percent in Class 13 to zero in Classes 9 and 10, with a median percentage of 11.8 percent...

Such a description is, of course, useful for some purposes, but it is not likely to capture the meaning or character of the teaching that has occurred. Such a description of behavior is 'thin' and can be contrasted to what Geertz refers to as 'thick' description. Thick description aims at describing the meaning or significance of behavior as it occurs in a cultural network saturated with meaning...

It is possible for critics to bring such bias to an encounter that they misread the situation. Their prior commitments function under exceptional circumstances as

blinders rather than guides for seeing what is happening. But this liability, too, is not absent from conventional research. Theoretical convictions can lead one to gross misinterpretations of classroom life and biases towards particular modes of statistical analysis or forms of testing can also create distortions in the state of affairs encountered. The tools we use are not simply neutral entities but have distinctive effects on the quality of our perceptions and upon our understanding...

One might well ask whether educational connoisseurship and criticism are likely to lead to useful generalizations about educational practice. Can the study of a handful of nonrandomly selected classrooms yield conclusions that apply to classrooms other than the ones studied? The answer to these questions is complex. Insofar as the application of critical procedures discloses subtle but important phenomena that other classrooms and teachers share, then of course the gist of critical disclosure is applicable. But the only way to know that is to be able to learn from critical discourse what might be worth looking for in other educational situations. In other words, if it is true that the universal does indeed reside in the particulars which artistic activity constructs, the renderings of those constructions in critical language should open up aspects of classroom life that participate in such universals. To know that requires itself a sense of connoisseurship. Unlike the automatic application of a standard, what one learns from effective criticism is both a content within a particular classroom and a refined sensibility concerning classrooms that is useful for studying other educational situations.

There is another way in which effective connoisseurship and criticism might yield warranted generalizations and that is as cues useful for locating phenomena that might be subsequently pursued through conventional educational research. Creative scientific work in any field depends upon new realizations, new models or new methods to guide inquiry. Insofar as effective criticism reveals aspects of educational phenomena that were previously unnoticed or underestimated, a fresh focus for conventional scientific study could be provided...

Excerpted from position paper

Elliot Eisner teaches art and education at Stanford.

How do you evaluate an eolith?

DAVID HAWKINS · In an essay published posthumously in 1952, the American engineer-novelist, Hans Otto Storm, undertakes to contrast two principles of human workmanship which he calls the principle of design and the principle of eolithism. The principle of design he describes out of the engineering textbooks and out of his own experience. He is thinking, for example, about something in the way of a bridge, a building, or a machine. The designer must first know approximately what he wants and how it is to be used. The next choice is a choice of ma-

terial from which to build, which must be of known and preferably uniform properties. This certainty and this uniformity are essential to the whole process; they affect not only the geometrical and physical result of good designing, they also affect deeply the mental discipline which the process of designing demands. Given a knowledge of the material and the final objective, the one is applied to the other and a plan begins to emerge, checked and extended by the use of well-known arithmetical rules; this process being continued until the whole becomes realistic in detail, making contact with the existing world. At this point the change is made from thought construction to physical construction. The direction of the operation is reversed, starting at contact with the material world and extending the structure until it embodies, finally, the objective with which the whole process started.

In this essay, Storm wishes to challenge an assumption which comes with the principle of design in our society, that this principle is basic and universal, an ideal by which we can measure all craftsmanship from its most primitive and blundering beginnings--a presupposition which all craftsmen are committed to, even if they are not trained or intelligent enough to understand the pure idea for which they should be striving.

To challenge this assumption, Storm, a professional radio engineer himself, puts forth an alternative and wholly different principle of workmanship, one as distinctively human, and I would say far more distinctively human than the principle of design. This is the principle for which he borrows the term of a slightly bastard etymology, 'eolithism'... An eolith is literally a piece of junk remaining from the Stone Age, often enough rescued from some ancient buried garbage heap. Storm quotes a definition, "stones, picked up and used by man, and even fashioned a little for his use." The important matter in the definition, Storm says, is that eoliths were *picked up*, already *accidentally* adapted to some end and, more importantly, strongly suggestive of the end...

(Now we come to the serious point of Storm's argument. To sharpen the contrast with the principle of design, he says, "let us remember the basic principle of the designing workman--he must know what he wants, and, even before design begins, he must decide on his material." The fashioner of eoliths, on the other hand, must have a continually open mind about materials, and he must be very adaptable in the matter of ends, of what he wants. If the eolith defies the use it first suggested, then perhaps there is another, equally interesting and worthy. The essential limitations of the principle of design lie in the givenness and fixity of goals, and in the need to eliminate variety and inhomogeneity from the means and materials, which are thereby reduced in any significance or value they may have *except* in serving those given ends...)

The principle of design, as Storm defines it, requires that we have goals well enough defined to provide criteria of choice among alternative means. It also re-

quires that we have materials available to us which are sufficiently homogeneous and sufficiently understood so that we can apply well-tested rules to the selection and organization of an efficient means of reaching our goal. Only when these two conditions are satisfied can we proceed to specify ahead of time just how our structure or mechanism will be put together.

The principle of eolithism, on the other hand, requires the nonsatisfaction of both the above conditions. To be effective, it requires that our system of resources be intrinsically varied and qualitatively rich, for that is the condition under which such a system, the world as immediately and concretely available to us, will be maximally suggestive of new goals. But such suggestions can gain a hearing only if established goals can be shaken and recast.

Clearly human life requires and exhibits an interplay of these two principles, which describe distinguishable but not separable phases of mind or--if you wish to avoid that sententious noun--of mindful activity. We seek goals and we set goals. As the seeking becomes routinized, we design mechanisms to help us. In seeking goals we encounter realities, however, which tempt and beckon, which bring us to redefine the goals we seek, and thus also to alter the directions of our seeking. The chemistry of these interactions allows a wide variety between extremes--those of monomania and disorganization. To understand this variety and to seek norms for guiding it is, I suppose, an explicit or implicit aim of philosophical ethics and, in particulr, of the philosophy of education...

The raw material which educational design would think of shaping not only lacks all the homogeneity of concrete and steel, but is inherently parceled in unique individuate form, in the form of live human beings. The product number, of possible congenital patterns times possible early biographies of children, is of a higher arithmetical order than the total number of children, past, present, or future. The probability is effectively zero that there should be two children presenting the same educational challenges and opportunities. One could, of course, assert the same heterogeneity of apples or blades of grass, looked at closely enough. But in education the heterogeneity of human kind is not trivial in that sense, because it affects the aims of the process, and is indeed required by them. It is a superordinate goal of education--and if it is not it ought to be--to help children on their way to become competent eolithic craftsmen in building their own lives, and this requires from the start a recognition of individual competency and situation. Not to recognize individuality is not to educate.

Thus, in our own designing for education and in the execution of such designs, there is required an essential component of Storm's other principle, of the eolithic mode. If we must talk about design, as we must, then let

it be design which facilitates and strengthens the capacities unique to our kind, capacities for rebuilding and reorganizing the specific commitments and pursuits, including education, that guide us and constitute our lives. To deny this component is to forget that the child becomes the adult and the pupil the educator--possibly even before the product has been duly graduated. We have excellent reminders, in our tradition, that children can be leaders as well as followers and it is even conceivable, at times, that universities could be instructed by *their* students... It is often conceded that a superordinate aim of education should be, what I have called it, the cultivation of competence in children to fashion well their own lives. But it is *not* supposed steadily that such competence is gained through exercise of it. It is supposed, rather, that self-organization will appear magically *after* years of schooling subordinated to a quite different principle, according to which children are deprived of autonomy...

As I see it, there is a new wave of educational design inimical, or at least indifferent, to the implications of the sort of critique I have been suggesting. Here I should like to speak more personally and more concretely. For two years, I was completely involved in an effort of elementary school curricular innovation. From the wide range of the natural sciences, there is a vast amount of concrete material, prized by scientists because of the intriguing and esthetically captivating phenomena associated with it, which are, at the same time, intimately interwoven with the deeper laws and histories of nature which science has found. To survey this vast field and take from it materials which can be put in children's hands, to entrap their interest and to stimulate their capacity for wonder and inquiry, seems an important undertaking for this age when science, so interwoven into our lives, yet remains for most an alien and forbidden territory. During the time I was involved with the Elementary Science Study, and in the work some of us have tried to continue here since, there have been many times when we were called on to do something called stating our objectives. Now that is always a reasonable-sounding sort of request, if made by someone who genuinely doesn't know what you are up to--and especially if you are asking him for financial support. But it gradually became clear to me that many people were not satisfied with the sort of answer I and my associates were prepared to give, and such as I have suggested above. I came to realize that it was somehow expected that we could lay down, in advance, a set of specifications as to what we hoped to accomplish, and by which our work could be something called *evaluated*. The truth of the matter was that we were explorers working our way through the heterogeneous world of schools, children, teachers, and phenomena of science. There are the resources; what can you make of them? In short, we were eolithicists being confronted to show cause why we should not be classed with the junkman and his customers. Somehow one was supposed to be

wise enough to define an all-encompassing set of ends before one had acquired the slightest bit of good sense about the nature and potentialities of the materials.

To all such questions, our answers were characteristically, and I must confess for my own part deliberately, coy and evasive. We often said some general sayings, hoping to suggest that we knew what we were about. We said we hoped to make it easier for teachers to induct themselves and children into a frame of mind conducive to the enjoyment and close observation of natural phenomena, and then into the practical art of scientific investigation; that, so far as we knew, this could only be done by getting involved in that art, from the beginning. This meant designing inexpensive laboratory materials and apparatus, and in the best eolithic fashion surveying the resources of wood and field and stream, of back alley and junk pile. We said that we therefore thought it best to try to *e*volve curricular materials and strategies out of repeated attempts to *in*volve children in inquiry, their inquiry. We were thus committed to be very opportunistic, that is to say very empirical, in selecting for further trial just those materials and strategies which did in fact best beckon to and absorb children of various ages and conditions. Nor did we believe that we could become final authorities on this subject. What we hoped, rather, was that in enlarging the store of materials and ideas available to teachers, we would help them in *their* proper task of helping children on the road to more competent choice and learning. That also meant giving teachers wider opportunities for choice and learning; not circumventing them with detailed curricular guidance which would substitute for their inventions and denigrate their professional role.

But often the demand for objectives was not satisfied with this kind of 'vague, loose talk.' What was expected often, whether from sheer habit or from anxiety or extreme narrowness of vision, was that we should produce a completely organized and sequenced guide for everybody who 'adopted' our program. And that was where we stuck. We said we thought we were learning some of the *means* of good science teaching, but that we were not yet nearly wise enough to present what is vulgarly known in the trade as a complete package, with objectives spelled out, little texts, and words in teachers' mouths. We said we thought this should be left quite flexible and open to decision in the careers of particular schools, teachers, and children; open to significant choice. And then, of course, came the clincher: "Ah, yes, but how can you evaluate your work if you don't state carefully defined objectives?" The phrase 'behavioral objectives' had not come upon the scene, final expression of the primate dominance order which puts 'design' on top, but we took the question to mean, as it typically does, a detailed setting forth of performance criteria related to subject matter and skill. And that, as I have said, we thought we should be more than a bit hesitant about. As I look back

on that period I think we should have been less hesitant. We should not have *answered* the question about detailed objectives of subject matter and skill, but we should have have explicitly *questioned* that question. Suarez said that it was the philosopher's job to question the questions, and answer the answers; to provide what a well-known mathematician, Stanislaw Ulam, once called "the necessary don't-know-how." So I, as a philosopher, especially should have been bolder...

Learning in an educationally important sense is an active process of self-organization and reorganization, which takes place through the mediation of choice among significant alternatives available to the learner. If you want to put this proposition in a general context, it is at the opposite pole from dominant theoretical positions associated in this country with the title "Learning Theory;" and in essential agreement with the theoretical position of Jean Piaget...

Excerpted from position paper

David Hawkins is professor of philosophy at the University of Colorado in Boulder and director, there, of the Mountain View Center for Environmental Studies.

Is accountability destroying education?

DAVID J. IRVINE: No one can argue with the principle of accountability. It is reasonable to expect that people who are given a public trust be held accountable. But how does the principle work in practice? It is time that the devices we have developed in the name of accountability were themselves subjected to accountability...

The emphasis on accountability has led us to think that absolute objectivity is required in studying education. 'Objective' tests are better than subjective tests. The teacher should view students 'objectively.' The evaluator should interpret data 'objectively.'

How would the Hitler youth or the murder of Jews be evaluated 'objectively'? If the programs were evaluated against their performance objectives, they were smashing successes. But is that kind of evaluation adequate? Our vantage point in history in relation to the Nazi regime allows us to say *No*. More basic values were involved. If that was true then, is it not also true now? The worth of program objectives must be judged before we evaluate the program's success in attaining them, and this is not likely to be an objective process...

Many of our present methods of accountability seem to be based on the implicit assumption that there is some 'best' or 'most valuable' learning. In a few areas involving the development of skills this may be true. But for many kinds of learning, Kipling's admonition is probably closer to the truth: "There are nine and sixty ways of constructing tribal lays, and every single one of them is right..."

Accountability may impede the development of meaningful objectives because those things which are easiest to measure may tend by default to become the major objectives of education. This might be called the 'Red Bird

Syndrome': The hawks are stealing our chickens, but we shoot at the red birds because they are easier to see in the underbrush.

Ironically, accountability, which is intended to enhance the quality of education, may tend to reduce the service which education provides society. One of the most common uses to which test results are put is to select students for the next level of education or for a particular program. This is viewed by many as a form of accountability, that is, as a process for maintianing quality. Yet what a perverse kind of 'quality' when it consists of keeping people *from* education!

Perhaps most important, accountability tends to focus our attention on that learning which has been important in the past. Preparing our children for their lives in an unknown future requires more than feeding them the learning from the past or even providing them with effective learning skills which they can apply in that future. The future will become what they can *make* it, and what they can make it will depend largely on the creativity, the resourcefulness, and the human qualities which they develop. Every element of the educational process--accountability included--must be bent toward these ends...

Excerpted from position paper

David Irvine is chief, Bureau of School Programs Evaluation, New York State Education Department.

A San Francisco parent's comments on evaluation:
'We must acknowledge our humility'

EFFIE KURILOFF: Evaluation always conjures a negative thing. I don't like being judged by other people. I don't mind being helped if I ask for it. I am constantly judging myself. I evaluate other people. I worked on a mural today. I was constantly solving problems about how to connect things, the goat, the rabbit, the trees. I spread the blueberry bushes, which was one of the children's ideas to connect things.

I resent evaluations in schools. It doesn't help. I don't like grades either. If I were in a group of people, I would ask for feedback from the people about myself or my idea. If I see behavior that I don't like (in children), I stop the behavior in a firm, but loving way. So I won't have to give an 'F' later for bad behavior. I hate punitiveness. I like educating, leading out from.

All of nature is perfect. The tree, the child, the ocean, the worm. And cannot be improved upon. The only thing we can do with it is appreciate it, enjoy it, respect it, love it, marvel at its perfection. As we enjoy and appreciate, we provide space for further growth and further appreciation. This can be translated into 'caring for.' Today on my way down to work on my mural at the recreation center, I picked a daisy and noticed all of its white petals circling the yellow center, and proceeded to paint a large bunch of daisies on the mural. Everybody who saw them just loved them.

So what if somebody decides that a student is a 'D'

or an 'A.' The teacher then treats him accordingly and the student behaves accordingly. He's boxed in. Both the teacher and the student have lost the opportunity to become, to play, to change; they are now a 'D' or an 'A.' Trapped. The possibilities for a person are limitless when he is not bound by someone else's limitations and expectations of what and who he is.

We must acknowledge our humility at the perfection that is a foot (I am massaging my daughter's foot while she writes down these random thoughts). You play with it, that's what you do.

How am I going to shape this and package it, this paper? And I realize I don't have to. I am not doing this to inform somebody else about evaluation. I am doing this to focus on the concept, to play with it.

All of life is practicing and gaining skills to live your life by and with. Becoming increasingly better breathers, noticers, painters, writers, bakers, foot massagers, mothers, gardeners, planters, and harvesters, nourishers, maintainers, so that each person can take care of himself--or herself--and care for his home, his family, his community.

If evaluating as it is practiced disempowers people, takes away their feeling of energy and strength, then we should get rid of it. For we cannot afford to waste anybody's strength. Indeed, we need each person working fully for himself, for his family, for his community. The process of evaluation must be empowering, enhancing, to everyone it touches, the evaluator, the evaluatee. Failing that we are wasting resources. We can no longer afford to waste. *Position statement*

Effie Kuriloff, a mother, and *David Schwarzchild*, a high school student, live in San Francisco.

The case study method in social inquiry

ROBERT STAKE· It is widely believed that case studies are useful in the study of human affairs because they are down-to-earth and attention-holding, but that they are not a suitable basis for generalization. In this paper, I claim that case studies will often be the preferred method of research because they may be epistemologically in harmony with the reader's experience and thus, to that person, a natural basis for generalization...

When explanation, propositional knowledge, and law are the aims of an inquiry, the case study method will often be at a disadvantage. When the aims are understanding, extension of experience, and increase in conviction in that which is known, the disadvantage may disappear.

Generalizations. The scientist and the humanist scholar alike search for laws that tell of order in their disciplines. But so do all other persons look for regularity and system in their experience. Predictable covariation is to be found in all phenomena. In 1620, Francis Bacon said:

> There are and can be only two ways of search and discovering truth. The one flies from the senses

> and particulars to the most general axioms...this is now the fashion. The other derives axioms from the senses and particulars, rising by a gradual and unbroken ascent, so that it arrives at the most general axioms last of all. This is the true way, but as yet untried.

He claimed that Truth lies in the most general of axioms, a far and labored trek from experience.

Another point of view holds that Truth lies in particulars. William Blake offered these intemperate words:

> To generalize is to be an idiot. To particularize is the alone distinction of merit. General knowledges are those that idiots possess.

Generalization may not be all that despicable, but particularization does deserve praise. To know particulars fleetingly, of course, is to know next to nothing. What becomes useful understanding is a full and thorough knowledge of the particular, recognizing it in new and foreign contests...

Sociologist Howard Becker spoke of an irreducible conflict between sociological perspective and the perspectime of everyday life. Which is superior? It depends on the circumstance, of course. For publishing in the sociological journals, the scientific perspective is better; but for reporting to lay audiences and for studying lay problems, the lay perspective will often be superior. And frequently that everyday-life perspective will be superior for discourse among scholars for they too often share more of ordinary experience among themselves than of special conceptualization. The special is often too special. It is foolish to presume that a more scholarly report will be the more effective...

Common characteristics of the case study are: that descriptions are complex, holistic, and involving a myriad of not highly isolated variables; that the data are likely to be gathered at least in part by personalistic observation; and that the study is likely to be reported in an informal, perhaps narrative, style, possibly using verbatim quotation, illustration, and even allusion and metaphor. Comparisons are implicit rather than explicit. Themes and hypotheses may be important, but they remain subordinate to the understanding of the case.

Although case studies have been used by anthropologists, psychoanalysts, and many others as a method of exploration preliminary to theory development, the characteristics of the method are more suited to expansionist than reductionist pursuits. Theory building is the search for essences, pervasive and determining ingredients, the makings of laws. The case study, however, proliferates rather than narrows. One is left with more to pay attention to rather than less. The case study attends to the idiosyncratic more than to the pervasive. The fact that it has been useful in theory building does not mean that that is its best use.

Its best use appears to me to be for adding to existing experience and humanistic understanding. Its characteristics match the 'readinesses' people have for added experience. As Von Wright and others stressed, intentionality and empathy are central to the comprehension of social problems, but so also is information that is holistic and episodic. The discourse of persons struggling to increase their understanding of social matters features and solicits these qualities. And these qualities match nicely the characteristics of the case study.

The study of human problems is the work of scientists, novelists, journalists, everybody of course--but especially historians. The historian Howard Butterfield (1951) recognized the centrality of experiential data and said:

> ...the only understanding we ever reach in history is but a refinement, more or less subtle and sensitive, of the difficult--and sometimes deceptive--process of imagining oneself in another person's place.

Case studies are likely to continue to be popular because of their style and to be useful for exploration for those who search for explanatory laws. And, moreover, because of the universality and importance of experiential understanding, and because of their compatability with such understanding, case studies can be expected to continue to have an epistemological advantage over other inquiry methods as a basis for naturalistic generalization. Unlike Bacon's 'true way' of discovering Truth, this method has been tried and found to be a direct and satisfying way of adding to experience and improving understanding.

Excerpted from position paper

Robert Stake is director, Center for Instructional Research and Curriculum Evaluation, University of Illinois at Urbana-Champaign.

Measuring the color and taste

MARVIN C. GENTZ·

> "Invest"ing our lives in activities
> and "Spend"ing our time
> and "calculat"ing our benefits;
> the "worth" of our loves and
> the "cost" of our hates and
> "add"ing to our experiential field and
> "subtract"ing from consciousness and
> "count"ing the risks we take and
> "multiply"ing our blessings and
> and and and and and

"measure"ing the color and taste
and smell and feel, and "far-outness" and
"close-in-ness" and good and bad
and high and low and
and and and and and and
and "work"ing,
and "produce"ing, "efficience"ing
and "effect"ing and "affect"ing
and "knowing for sure
forever

is to hang parts of people on
the numerals ("by the numbers?", "by the book?",
"by their thing(s)?",
by the and, and and?) of "time-pieces," "Big
Bens", "clocks"
"dip-sticks", "guages", and and,

makes us all clock-watchers,
waiting for the bell to ring

the end.

Position statement

Marvin Gonta is an evaluator in the Experimental Schools Project, Berkeley.

3

Evaluation and Public Policy

On understanding cultural differences

ELLIOT EISNER· Because I am a human being, I can understand the experiences of many people who are not myself. I think there's an enormous amount of testimony to that. I think it's represented in art; I think it's represented in our ability to understand each other. I just don't accept the idea that I have to be of a particular race or class to understand people of that race or class. If that argument holds, what guarantees my understanding of that individual's experience if he's sitting up at this table? If I can't understand a black man, red man or Chicano sitting in the audience, coming up to this table isn't going to make a difference.

MARIA VARGAS· I disagree. I have been working with people who have different experiences. By the way, I went to a regular Ivy League college, so we share a lot of experiences. But there are cultural differences in those experiences and I can't pretend that I am going to know what living in a white suburban community is like or a black ghetto. In teaching, people who share the culture tend to relate to their own people better. And it's best to let them decide for themselves what is important.

Exchange during panel discussion

White evaluators and black evaluatees: a dialectical discussion

JAMES A JOHNSON, JR.· *Pro:* You've charged that I have confused the role of the decision-maker and the evaluator. The fact of the matter is that the evaluator *is* a decision-maker. The evaluator decides, among other things, what questions will be asked, how the evaluation problem is to be defined, how the problem and evaluative questions are to be operationalized, what instruments will be administered, what population will be used, how the data will be treated, and how they will be interpreted. Second, I am not suggesting that the average evaluator behave in a manner that is inconsistent with the culture by which he or she is bound. But, at the same time, I am suggesting that the average evaluator is less qualified to evaluate educational programs in which persons not of his or her

culture or life style are clients than programs in which the experience bases of the evaluator and client are consistent. Third, theory, concepts, and technologies are cultural artifacts. Deprivation theory leads to deprivation programs but stems from obtuse and ethnocentric perceptions of the poor. The behavior of poor persons is governed by rules that are different from the rules that govern the behavior of the affluent. Yet the rules of the affluent are imposed on the poor. Educational need, then, is a function of the extent to which the behavior of poor persons deviates from the alleged behavior of the affluent. With respect to concepts, I cannot conceive of an instance in which poor and affluent persons would hold similar concepts of density or hunger or wealth or happiness, much less color, shape, or size. If a Navajo became ill and had to be admitted to a hospital, it is highly probable that no six-sided room would be available. Phenomena such as time, space, territory, and temperature, and coexistence are culture-bound. An evaluator, good or bad, should not be a mannequin. That is, evaluators can no longer be permitted to use available technology to generate numbers that answer questions posed by the evaluator but are irrelevant to the needs of the client. You speak of 'competent' evaluators. Well, I would bet that, as a rule, the more 'competent' the evaluator, the more likely it is that the evaluator will have a frame of reference that is male, white, and middle class. What would he know about the poor? How would he perceive women? What are his expectations of persons of color?

Con: If you are suggesting that in order for a program to be adequately evaluated there must be a one-to-one correspondence between the evaluator and benefactor, I reject your argument out of hand. The competent evaluator is obviously a highly trained individual. He or she has had years of experience in the field, established a good track record, and draws conclusions about a prgram, process, or product that are supported by evidence. The evaluator does not make programmatic decisions. To the contrary, the evaluator merely provides decision-makers with evidence by which they may make more rational decisions than they would were data not available. Again, I say any well-trained, competent, and experienced evaluator would be able to assess the value of programs, processes, and products regardless of the user or benefactor.

Pro: And what form would that assessment take?

Con: Well, for example, the evaluator may conclude that the program or process or product does or does not behave at criteria.

Pro: You mean that if the evaluator asks the question: "Does it work," the evaluator can collect evidence which would yield a yes or no answer?"

Con: Yes. If one is evaluating the effects of an

intervention on student achievement, for example, one could test students who received the treatment and students that did not; compare...

Pro: But, how would you know that the two groups were comparable in all respects with the exception of the treatment?

Con: You assume that the groups are comparable by controlling for such variables as age, grade level, SES, mother's education, sex, ethnicity, etc.

Pro: But even after you match the groups on those variables, how do you know that the groups are comparable? In other words, if the students in the treatment and control groups are not assigned randomly, how does the evaluator know that effects are a function of the treatment and not a function of a systemic difference between the control and treatment groups?

Con: If the groups are matched on an adequate set of variables, it is assumed that...

Pro: Suppose you have an equal number of boys and girls in both groups. All are 9 years old, all are from low-income black families, and none of the mothers finished high school. But, because of some implicit systemic school practice most of your control group students are from 'getting ahead' families that attend 'the church of what's happening now' and most of your experimental group children are from 'doing all right' families that attend a traditional Baptist church.

Con: The groups would obviously be incomparable.

Pro: How would your well-[illegible]ned, experienced, and competent evaluator know? Do you know about 'the church of what's happening now'? Are you aware of the 'doing all right vs. getting ahead' orientation among low-income black families? Have you ever controlled for these variables?...

Con: You don't understand. I am talking about conventions. I'm talking about years of experience with a technology that has some problems but is the best thing available...

Pro: But that is not good enough. What you are doing is imposing a model that may be consistent with middle class rule systems on a population for which that model was not developed. For about fifteen years now, we have had politico-Ed-Psych types, among others, playing well-remunerated rubber-chicken circuits, proclaiming that the quality of schooling is not at issue, but rather it is the quality of mothering that contributes significantly to academic achievement in schools. There is an Ed-

Psych type and a guy who makes transistor radios who would explain differential achievement genetically. Many of these people are today arguing that increasing resources for the education of poor black children makes minimal difference. Yet, the only thing that you can tell me is that "we've been doing it this way for a long time" and "it's the best thing that we have available."

Con: I understand how *you people* feel. I myself come from a poor...

Pro: It may have been poor but it wasn't black. If it were, there is an extremely low probability that you would be at this conference. Look around you. How many black *evaluators* do you see? I am saying to you that there are fundamental differences between the experience bases of poor and affluent children. Poor black children learn their mother tongue via a process that in no way resembles the process by which white and middle class children learn language. Poor black children play and learn in multi-aged peer groups (sometimes ranging from age 2 1/2 to 13). How does your model prevent radiation effects? How do you know that your treatment groups are in fact not receiving the treatment vicariously? Have you as an evaluator ever asked the question, "Does the program have legitimacy?" or "Do the clients and benefactors perceive the program's projected outcomes to be in their best interest?"

Con: Actually, I have been more interested in decision evaluation -- evaluation which results in public policy recommendations. You know, cost effectiveness, cost benefits, educational productivity...that sort of thing. I can't get into the soft stuff--like cultural rule systems, cultural mismatch, and incongruity between home and school, child-expectations and incentive systems. I wouldn't know where to begin...

Pro: I disagree. You could get into it if you had to. If black people said Stop!!! We have had enough of this. You are not collecting any more data on our kids until it can be demonstrated that you or a senior member of your evaluation team can demonstrate that the sociocultural context in which our children are reared is understood, and until the evaluative process is consistent with and reflective of that process.

Con: Well, you don't have to get hostile about it. My goodness!!!

Pro: I don't view my position as hostile. What you must understand is that I am well aware that your theory and practice are consistent and supportive of a program that is in your best interest--intellectually, financially and politically. I am suggesting that the program cannot be in my interest. It does not have legitimacy, in my

view. It works, but it works against my interest.

Con: Then I guess that we have joined a very fundamental issue.

Pro: I guess that we have. I cannot, after all, be expected to make a total commitment to a program that is not in the interest of poor black children and youth. Yet you must remember that to the extent that poor black subjects dry up, that is the extent to which your program is inoperable. Furthermore, on moral, philosophical, and professional grounds, my perspective should be judged to be co-equal with yours and, therefore, worthy of debate.

Con: I will admit that you have raised one or two points that deserve further discussion. I don't know what the answer is. We have tried to find qualified black evaluators, but these efforts have not been very fruitful. When, on occasion, we have attracted one, someone inveigles him or her to leave for more money. I just don't know what the answer is...

James Johnson is executive director of the Jefferson County Education Consortium, Louisville, Kentucky.

Pro: What is the question, and what are your alternative courses of action?

Excerpted from position paper

Education, life chances, and the courts: the role of social science evidence

HENRY M. LEVIN: What is the proper role of social science in charting educational policy for improving the life chances of low-income and minority students? The answer to that question is not clear. The question of the relationship between educational influences and actual adult status addresses a very complicated area of social and individual behavior. In particular, little is known about the effects of different school environments on human behavior, about underlying theories of human productivity and its determinants in a particular social setting, about the myriad of other influences that can intervene between the educational strategy and the adult outcomes many years hence, and about the appropriate measurements of even those factors that do seem relevant. Further, the fact that experimentation as an empirical investigating tool is politically and practically infeasible limits severely our ability to uncover the true relationships.

Some observers may react to these conclusions by suggesting the social science evidence in these complicated areas is likely to be so misleading and value-laden that we ought to ignore it. In contrast, some technocrats will argue that the case *against* the ability of social science to validate the relationships between education and life chances has been overstated and that rapid scientific

advances in research methodology will even nullify those anomalies which have arisen. Both of these views assume that the social sciences must play a deterministic role in contributing to policy or that they can play no role at all. Yet, it may be the heuristic aspects of social science research which are most useful.

Alice Rivlin has suggested that we acknowledge the development of a 'forensic social science,' rather than pretending to be part of the tradition of balanced, objective social science in which the scholar hides (or claims to hide) his personal biases, and attempts to present all the evidence on both sides of a set of questions so that the reader may judge for himself. Using the notion of a forensic social science for addressing policy issues,

> Scholars or teams of scholars take on the task of writing briefs for or against particular policy positions. They state what the position is and bring together all the evidence that supports their side of the argument, leaving to the brief writers of the other side the job of picking apart the case that has been presented and detailing the counter evidence.

The problem with such an approach is that it assumes that all of the sides will be fairly represented. But adversary proceedings normally are based upon only two conflicting points of view. Moreover, the fact that the epistemology of the social sciences itself limits the analysis to a specific set of hypotheses (particularly ones that have readily identifiable empirical consequences) suggests that the issue might be framed in an erroneous manner. Of course, this type of bias can be avoided by permitting nonsocial scientists to enter the forum to present their views and argue their evidence. It is not clear what criteria would be used to select such witnesses nor is it obvious how one could determine how many points of view should be permitted. It is also not clear that the courts would attach great weight to 'nonscientific' presentations that are grounded in complex statistical procedures and highly technical language. Of course, the court could hire its own experts for examining and interpreting the evidence, but what guarantees the objectivity of the 'wise men' who advise the court?

Social science research can best be used to frame the issues and their consequences rather than to obtain conclusive evidence on what is right and what is to be done. This approach requires a recognition that while many aspects of the world cannot be quantified or analyzed in a social science setting, such factors should be considered along with the results of social science research. It is not clear that utilization of social science research in this manner is consistent with an adversary framework. Further, if social science findings increasingly are used to create what appear to be technical is-

sues out of essentially moral dilemmas, this presents a potential social danger. The apparently increasing reliance of the courts on social science evidence suggests that intensive debate on these issues should be given high priority. *Abstract of position paper*

Henry Levin teaches education at Stanford.

On values and public policy

ROBERT STAKE · It seems to me there are among us people who have a strong hope that through enlightenment, understanding, seeing the situation better, we will be able to work together on the state, federal, district, and classroom levels and come up with an evaluation system that's better than the one we have got and takes care of our problems. My feeling is there is no way to satisfy needs for information for people who are a long way from the classrooms. And that through some kind of adversarial adjustment to people screaming loud about the indignities of evaluation, some adjustment may occur; but it will be at the expense of any information that flows to district authorities.

ELLIOT EISNER · This conference represents an indication of the kind of irritation that exists in this country. What we need is continued clarity with respect to our aspirations. The gap between irritation and aspiration will be closed. I think we are going to be able to realize some of the aspirations that are beginning to emerge.

WALTER DENHAM · Bob mentioned at the beginning about information being available and what use it should be put to. I would speak to the point of decision-making and dollar allocations--somebody or some group will allocate dollars from time to time and will make judgments about whether they are satisfied or not with the way the dollars have been allocated. In terms of public policy, I think the issue is what information will these policy-makers and budget-makers have. I think it may be a myth that the only data they will accept is standardized test data. But there is power, there is control, there are budget decisions and there will continue to be judgments made, etc. I think the right issue is to ask what information should be provided and maybe 'should' is a useless word anyway. What information will be accepted or what alternative sets of information can be provided and potentially accepted? I think the adversary role where people don't want to be evaluated or want to lobby for their own point of view or their own way of spending money is not likely to be successful. There must be some kind of information provided, and I would rather not use the word evaluation in talking about what information is to be provided, because I think it tends to set people up in adversary or argumentative modes right away, depending upon their own definitions of the word 'evaluation.'

Walter Denham is director of program planning, Elementary Field Service Unit, California State Department of Education, Sacramento.

MARIA VARGAS · Something that bothers me a lot is that we are not providing information for people who are objective --we are providing information for people whose values aren't being examined, people who are racist, classist, and sexist. That is yet to come up in any discussion I have heard. I am very concerned about that. I'm very concerned about allowing people to determine what in their culture will be important to them. It is really important to consider that at every level. We have people, who never question their own points of view, making decisions, and I would really like to challenge that and then maybe I will consider the kind of information I will give them.

BILL SPADY · I see the topics of discussion in the conference revolving around three broad domains. One has very much to do with values, and the awareness that everyone here has his own values, biases, and windows to the world that tell what is important to education. The second is technological--i.e., building the tools to tap into some of the demands that a majority of people here are making. The third is the political or power domain. The way they hang together is that the values and awareness domain really influences the shaping of the technologies in what I would like to call research rather than evaluation; and the values and awareness domain also then shapes the nature of the decision-making, policy-making, and power struggles that go on. What I hear the conference participants being concerned about is finding a way to take their value perspectives and their awareness and translate them into both philosophies and technologies about the use of evaluation and research that will somehow impact on the policy-makers.

It's my thesis that there are three primary levels of government that we often confuse. The local school board is an enemy because it wants the local district to evaluate things in a certain way. The state--both the Department of Education and perhaps the State Senate, perhaps the State Legislature, perhaps the State Board of Education--is the enemy because there are mandates for doing things in a certain way coming down from State and we also know that the various components of State don't agree with each other. The federal government--where I am situated--we are the enemy because, when we do something like an experimental schools program in Berkeley or we fund other kinds of experimental endeavors here and there and people want evaluations, then you have to contend with those things.

My basic feeling is that the struggle between class and values that Maria Vargas just talked about is a struggle between the dominant value orientation of American society, which is lower middle class, and other class groups in the society; and that school as usual--good old regular school--is dominated by an ethos that emanates out of and is strongly influenced by lower middle class values which are different from upper middle class values

and other class values. The emphasis on standardized testing--knowing the kids can read and write, so that they get ahead, are upwardly mobile, get a good job, make over $15,000 a year if possible--is a very, very strongly lower middle class orientation. Who composes local school boards? Largely lower middle class people. To what constituency do state representatives respond more than to anything else? This dominant lower middle class constituency. What does Congress respond to? The same constituency. So the arguments I hear here about a new value system about evaluation and research that gives us a new approach to technology in evaluation and research and that would hopefully lead to the shaping of a new set of policies about evaluation and research is a conflict that emanates very strongly out of the dominance of a lower middle class ethos in this country and the values and priorities that other social class groups have about what is important in education.

Bill Spady is on the research staff of the National Institute of Education, Washington, D.C.

ALEX PAPILLON · I heard one person talk about lower white middle class values dominating the educational system without mentioning that the dominant ethos in that area is to aspire to be upper middle or upper class. Which bothers me, because all he is really saying is that the dominant value is upper class but the people who control the situation and the system are those who are aspiring to be part of that class.

It also bothers me to hear someone talk about the technology of evaluation vs. the decision-making process of policy decision-makers as if they are distinct entities. That kind of microcosmic look at reality is no more than a reflection of the practices within the technology as opposed to the realities the children live in. Whatever the analytical decisions that the technocrats and policy-makers made, I as a black parent and as a black student still came up with the same results. Now someone said we have to make a division between the testing by the teacher and the evaluation mechanism. That amazed me because the teacher called me dumb and the evaluation mechanism called me dumb and I have to figure out how they came to the same conclusion when they are supposed to be different.

One other point--the evaluation mechanism says that third world children are failing in the system. Teachers actually fail third world children. Policy-makers actually decided that third world children are not the priority, their needs are not the problem and they present a problem to society. It amazes me how the three distinct compartments have arrived at the same conclusion and third world children wind up being the victims of all of them. Stop being so parochial in your approach to human problems that have been affected by certain people being in decision-making positions vis-a-vis the evaluators, teachers, policy makers. The reality is that each one of these levels has produced the same results, and the same victims out of each one of those individualized processes.

I think we may have to talk about the whole process as opposed to indoctrinating ourselves with the idea of being the benevolent solution-giver. This goes back to the issue that a brother raised: you think you may be the answer and in fact you are part of the problem.

Exchange during Conference panel discussion

Testing and tracking

KENNETH KENISTON · Most of the fundamental objectives of Head Start seemed to be overlooked by the critics when the program was being evaluated. As conceived, the program was intended to give power to parents, to broaden children's experience in noncognitive ways, and to provide services such as health and dental care.

Yet most critics of Head Start seemed to hold that the program should have been able, in a few hours each day, to overcome permanently the overwhelming disadvantages of children born into poverty and discrimination. It is a tribute to our optimism, if to nothing else, that we ever thought so little could do so much for so many.

The theory underlying much of the evaluation of Head Start attributed the plight of those children to 'cultural deprivation,' meaning chiefly lack of intellectual stimulation. Yet this theory leaves us all blind to far more fundamental things. We need to get at what is causing the cultural deprivation--what is doing the depriving.

It seems clear that the reason many families cannot provide their children with intellectual stimulation for breakfast and cultural enlargement for dinner is not lack of culture but lack of money--because they are bogged down in the morass of old-fashioned poverty. Poverty is a manifestation not of our cultural system but of our economic system.

I have emphasized Head Start to underscore our national obsession with cognitive development as measured by test scores. We tend to rank and rate children, to reward and stigmatize them according to their ability to do well in the narrow tasks that schools (or we psychologists) can measure quantitatively.

This same ability to do well on tests is a primary determinant of the child's progress and position in the world of school and, to a large degree, in the later world of adults.

Why is this? Once again, I would not blame teachers or parents, but would point to the pressures of an advanced industrial economy. In our highly developed technological society, we have adopted, usually without knowing it, the implicit ideology called 'technism', which places central value on what can be measured with numbers, assigns numbers to what cannot be measured, and redefines everything else as self-expression or entertainment. 'Objective' measures of I.Q. and performance (which are, in fact, not at all objective) are an expression of this

broader propensity.

Thus, we measure the effectiveness of education by whether or not it produces income increments, not by whether it improves the quality of life of those who are educated. And we measure the success of schools not by the kinds of human beings they promote but by whatever increases in reading scores they chalk up. We have allowed quantitative standards, so central to our adult economic system and our way of thinking about it, to become the principal yardstick for our definition of our children's worth.

A related characteristic of our technological economy is its need for some mechanism to sort individuals into various occupational slots. The intellectualization of children by testing and tracking in schools assists in classifying and sorting them for the labor force. By the time a poor, black, handicapped, or uncared-for child reaches fourth or fifth grade, a consistent position in the bottom track of the grade has become an almost inescapable adult destiny. Hence, although we talk about the other human qualities we cherish in children, when push comes to shove--when it is a matter of promotions, credentials, and praise--the child who has learned to master test-taking gets the reward.

And this fact lives next door to our professed devotion to other human qualities we say we value more--physical vitality, caring, imagination, resourcefulness, cooperation, and moral commitment.

Excerpted from position paper

Kenneth Keniston is professor of child development at Massachusetts Institute of Technology.

The invention of failure

LEE CONWAY · For thousands of years lineage served as the impeccable standard, the stable criterion all men could look to for recognition of their place and that of others in the scheme of things. Once lineage became obsolete as a method of identifying and ranking individuals, new, rational criteria for evaluating people had to be established. These new criteria had to be objective, in keeping with a growing emphasis on rationality and efficiency, implicit in the needs of technologically advancing societies. As the value of lineage and family origin decreased, the new criterion became individual achievement.

Although 'objective' evaluation of achievement has replaced evaluation according to bloodline, the end result has not been a freeing of individuals to fully realize their potential. The problem is that the identification of people as units of achievement has been coupled with the invention of the notion of failure. Our fiscal and social economies are both based on scarcity values. Thus, there is supposedly only so much wealth and only some of us can have it. The rest must remain poor. There is only so much achievement, and only some of us can have it. The rest are expected to fail. Whereas, in traditional society, one's intrinsic lifelong worth (or

lack thereof) was generally established by one's family, today one's identity is constantly under evaluation and the risk of being discredited is an omnipresent hazard. Hyper-evaluation, the process which guarantees that sufficient individuals will fail, thereby removing them from the race quietly and without bloodshed, nicely supports a scarcity-based societal structure.

Hyper-evaluation exerts its greatest destructive power during the crucial intermediate and junior high school years when it takes huge steps to eliminate entire ethnic groups and income brackets from the competition for societal benefits. The power to evaluate is the power to destroy; and educators have been invested with the tools to irrevocably ruin the life chances of up to 50 percent of their students. These tools are called standardized tests...

Nationally, children from black, Chicano, and poor white families average two years below grade level in reading and mathematics. Teachers and counselors throw up their hands when they see such scores and declaim, "what's the use."

These marks: IQ, grade level, achievement percentile, aptitude level, are derived from standardized intelligence tests, achievement and aptitude tests administered to all children from kindergarten through high school. These marks are used to intimidate students, convince their parents that their children are dumb, and provide the educational system with an 'objective' defense of its own inadequacy. For fully 50 percent of all school children, these scores constitute marks of oppression.

All standardized tests are formed (constructed) so that half of the pupil population are ranked above the 50th percentile, the mid-point; and the other half below the mid-point or median. Thus, by definition, one half of all children taking a standardized test must score in the bottom two quartiles, the lower half. These tests are so designed as to automatically create as a class of instant losers fully 50 percent of all children in America...

Beginning with kindergarten, the scores from these tests are included in a child's permanent record file, called a cumulative folder. These 'cum folders' are avidly read by successive instructors, becoming the child's silent stalkers, his critical autobiography. Cum folder information is usually a superficial, if not false, portrayal of the child; and frequently misused by uncaring teachers eager to get an easy 'fix' on the child who displeases them.

Disadvantaged youths, with their accumulation of failed subjects and substandard test scores, have proven beyond a doubt--based on 'scientific' testing, based on educational psychology, and based on modern statistics--that they haven't a chance in hell. The marks in their cum folders set them off as 'losers,' deficient in scholastic aptitude, general intelligence, and cognitive achievement. On paper, these kids didn't have the smarts

to make it. And, officially, test scores are considered inviolate evidence of achievement, and consequently ability and merit, in the public schools of America...

Evaluation, with its almost exclusive reliance on standardized tests, as a direct, causative factor in educational failure has been little investigated. In a 1974 study which I completed for the National Institute of Education of the U.S. Office of Education, it was determined that current testing modes are inimical to the self-image and motivation of black and Chicano children, shattering their self-confidence, lowering teacher and parental opinion, reinforcing negative expectations, and possibly contributing as much to the child's overall socio-educational deprivation as any other known factor.

Structured observations were conducted in the classrooms of seven elementary schools in the low-income sections of one of the 10 largest districts in California. Observations during standardized achievement testing determined that group achievement testing is a high-stress activity for a great percentage of low-income area children.

Indications of this stress and tension ranged from deliberate distraction of other children to crying and running out of the examination room. Additional incidents illustrative of disassociation, fear, and stress included the following: boys playing tic-tac-toe under their table several minutes after the test had begun; boys noisily pushing carrels together despite teacher warning; children insisting they had stomachaches; pupils with 'killing' toothaches; epidemics of bubble gum blowing, etc. Several intermediate grade children vomited on their test booklets.

Signs of frustration, conflict, and hostility were ubiquitous in all classrooms. A large percentage of children played it 'cool,' hurrying through the test; randomly guessing, then turning in their test and rushing out of the room. The average student was timed in 18 minutes for test batteries which, properly taken, require 38 minutes...

In Hellenic Greece, the contest was the essential mechanism of social mobility as well as the prime method of distributing prestige and social status in a manner acceptable to the free citizen. The chief moral values of Greece were represented in the contest. Similarly, in contemporary Western society, the standardized test is the technological analogue of our dominant morality.

Under the clever guidance of an influential educational psychologist named Edward Thorndike, Western education forged the quintessential tool to emulate the Hellenic Greeks. Thorndike, the godfather of standardized testing, dominated 'scientific' education between 1900 and 1940. His theories and his testing prototypes control education to the present day in America.

Edward Thorndike believed that character, goodwill, and cleanliness objectively correlated with intelligence. And by the time he and the army of educational psychologists he had trained were finished, by God, they did cor-

relate. Thorndike provided the first 'scientific' tests utilized to classify students. Clarence Karrier so aptly exposed the game, stating: "The social class system which emerged from his (Thorndike's) statistical charts reflected, he thought, a natural hierarchy of native intelligence."

Thorndike and his disciples fashioned their so-called intelligence tests into powerful standardizing instruments with which to limit and control mobility on the educational frontier. Then, with the help of educational psychologists, statisticians, and counselors, American educators could proceed to segregate those said to have superior intellect from their less fortunate fellows within the common public school. All of this deception could be done under the guise of concern for individual differences among children, the new science of education, and the 'objective' classification of students.

Thorndike's goals were to identify and set apart an elite of achievement, selected, shaped, and standardized according to middle-class values. He was extraordinarily successful. Contemporary tests are so constructed that a child can perform well only by being reared in a middle-class environment, since a good number of the concepts and items included in standardized tests are based on middle-class objects and values.

At the macrosociological level, tests can be perceived as clever instruments for gaining acquiescence to the early selection of an elite and ejection of the 'masses' from competition. As Edgar Friedenberg, among others, notes, test data are needed to justify decisions for which no data were needed at all, in order to make a show that the educational process if fair, impartial, and plays by the rules.

Turner, a sociologist, states that "system control is maintained by training the 'masses' to regard themselves as relatively incompetent...by cultivating belief in the superior competence of the elite. The earlier that the selection of the elite recruits is made, the sooner others can be taught to accept their inferiority and to make 'realistic' rather than phantasy plans. Early selection prevents raising the hopes of large numbers of people."

The standardized test nicely serves the purpose of maintaining scarcity by convincing a sufficiently large segment of the population that they 'blew it,' that the elite of achievement deserves their success. The many slots reserved for the failures then can be regularly and routinely filled with uncomplaining victims, all subscribing to the well-popularized, dominant morality.

Thus have the schools, in alliance with psychologists and test publishers, succeeded magnificently in the task of constructing an evaluation system which provides a living for educators while maintaining the status quo by generating failure. The evaluation system guarantees that the child will come to expect, then apathetically accept, defeat and failure, all the while believing the system which 'creamed' him was just.

Excerpted from position paper

Lee Conmy is an evaluation consultant in Richmond, California.

Also available as part of the North Dakota Study Group on Evaluation series:

Observation and Description: An Alternative Methodology for the Investigation of Human Phenomena
Patricia F. Carini

Alternative Evaluation Research Paradigm
Michael Quinn Patton

An Open Education Perspective on Evaluation
George E. Hein

A Handbook on Documentation
Brenda Engel

The Teacher Curriculum Work Center: A Descriptive Study
Sharon Feiman

The Word and the Thing: Ways of Seeing the Teacher
Ann Cook and Herb Mack

Testing and the Testing Industry: A Third View
John Williams

Special Education: The Meeting of Differences
Steven D. Harlow

Psychological Effects of Open Classroom Teaching on Primary School Children: A Review of the Research
Robert A. Horwitz

Children's Interactions in Traditional and Nontraditional Classrooms
Sylvia Ross, Herbert Zimiles, and David Gerstein

Developing Hypotheses About Classrooms from Teachers Practical Constructs
John Elliott

A View of Power: Four Essays on The National Assessment of Educational Progress
Paul Olson

Single copies $2, from Vito Perrone, CTL
U. of North Dakota, Grand Forks, N.D. 58202